Characterization of Novel Genetic Alterations in Prostate Cancer

Inês Teles Siefers Alves

The research described in this thesis was conducted at the Department of Pathology and Department of Urology from the Erasmus Medical Center and according to the requirements of the Molecular Medicine Post Graduate School (Erasmus MC).

The research in this thesis was financially supported by the FP7 Marie Curie Initial Training Network PRO-NEST (grant number 238278).

ISBN: 978-94-6259-774-7 (printed version)

Layout and printing: ΔIO proefschrift, www.proefschrift-aio.nl

Characterization of Novel Genetic Alterations in Prostate Cancer

Karakteriseren van Nieuwe Genetische Veranderingen in Prostaatkanker

Proefschrift

Ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
dinsdag 08 september 2015 om 13:30 uur

door

Inês Teles Siefers Alves
geboren te Lissabon

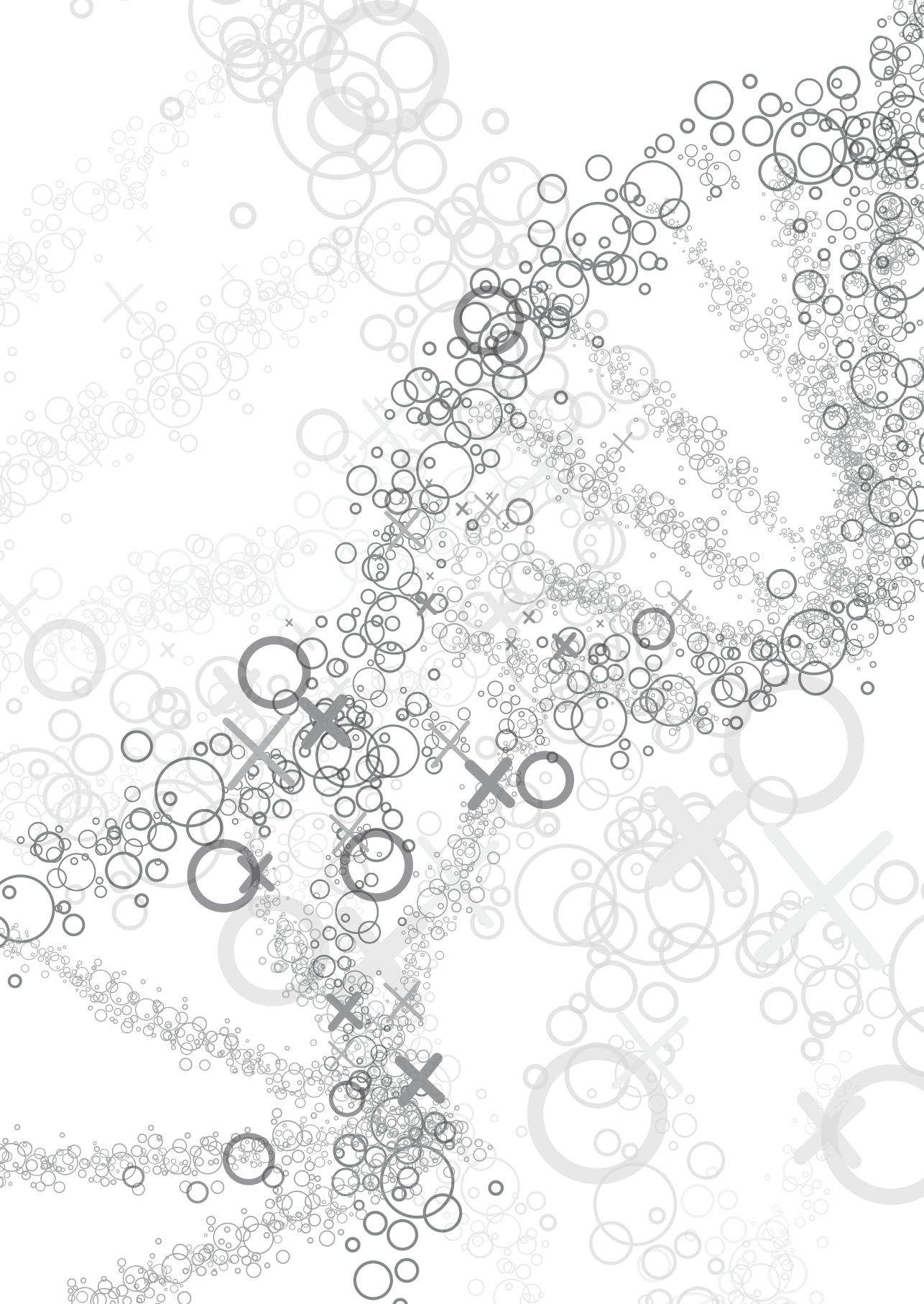
Erasmus University Rotterdam

Promotiecommissie

Promotor:
Prof.dr.ir. J. Trapman
Prof.dr.ir. G. Jenster

Overige leden:
Prof.dr. P.M.J.J. Berns
Prof.dr. A.G. Uitterlinden
Dr. G. van der Pluijm

Table of Contents:



Chapter 1

General Introduction and scope of the thesis

Molecular biology of prostate cancer

Ines Teles Alves[1,2]
Jan Trapman[2]
Guido Jenster[1]

Departments of Urology[1] and Pathology[2], Erasmus MC, Rotterdam, The Netherlands

Adapted from: The Oxford Textbook of Urological Surgery: Chapter 6.2 - Molecular biology of prostate cancer. Editors: Professor Freddie Hamdy and Mr Ian Eardley.

In Press 2015

Content:

Abstract

Prostate cancer is an heterogeneous disease that arises through the acquisition of key malignant hallmarks. At the molecular level, prostate tumours are dependent upon the androgen receptor pathway, which affects cell function, growth and behaviour through downstream androgen regulated genes. Prostate cancers require this activity and manipulate the AR pathway to maintain signalling. For example, mutation of the AR (to bind ligands other than androgens) or amplification/duplication of the AR allows signalling to continue in castration levels of testosterone. Around 50% of prostate cancers have a gene fusion between the androgen regulated *TMPRSS2* gene and a transcription factor of the ETS family, in particular *ERG*. This results in aberrant androgen stimulated tumor growth and survival. Current research is using molecular knowledge to identify biomarkers and develop new therapies for castration resistant disease, such as enzalutamide and abiraterone acetate. Along with the advances in next generation sequencing technologies, the complex landscape of genomic abnormalities and novel genetic phenomena like chromothripsis and hypermutation in prostate cancer have been revealed.

I. Molecular basis of cancer

The functional unit of life is the cell and the genetic information is transferred to daughter cells via its DNA. This is also true in tumours, which derive from an abnormal cell by accumulated DNA mutations, transferring these mutations to progeny cells. Cancer is a disease of DNA and RNA. Each human cell contains about two times 3.2 billion base pairs, divided over the 22 chromosome pairs and two X chromosomes in females or X, Y in males. About 30,000 known genes encoded by the human genome are responsible for the various cellular functions, mainly executed by the proteins encoded by the ~21,000 protein coding genes (1). More than 40% of the genome consists of genes, but only ~1.8% of the genome (the exons) will be transcribed to mRNAs and subsequently translated to proteins. It is unlikely that a random single change in a base pair will have an effect on the properties of the host cell. The chance of effective mutations in the right combination of genes needed for tumour initiation is extremely small. However, with a life-long accumulation of DNA damage and 10-50 billion cells in our body, in the long term the development of a tumour seems inevitable.

There are various types of DNA changes (mutations) that affect the function or expression of genes:

- *Deletion*. Loss of one copy of the DNA (in one copy of a chromosome pair) is referred to as loss of heterozygosity (heterozygosity is the presence of two different copies of a genomic region in the two chromosomes). Complete loss of two copies of a specific chromosomal region is called a homozygous deletion.
- *Amplifications.* This can vary between short or long duplications of a chromosomal region to tens of additional copies (high level amplification).
- *Translocations* refer to exchange of DNA fragments within or between chromosomes.
- *Insertions.* Additional base pairs can be inserted due to replication mistakes, but also viral DNA or genomic integration of amplified DNA are covered by this term.
- *Base pair changes* are defined when one or more bases are different without loss or gain. If the same base pair change is observed in >1% among normal healthy individuals, it is called a common variant single nucleotide polymorphism (SNP). If it is less frequent it is named a rare variant. The change is called a mutation if it is unique for the genome of a tumour, and not present in the DNA from healthy cells of the same individual.

The type of DNA change can have specific effects on the properties of the mutated genes. The genes that cause cancer are classically divided into two types:

- *Oncogenes* are genes generally encoding tumour growth stimulating proteins. A single mutation in one gene copy is enough to activate its oncogenic function. Amplifications are typically associated with oncogenes, but also base pair changes and translocations can cause activated oncogenes or oncogenic fusion genes.
- *Tumour suppressor genes* generally encode proteins that inhibit tumour growth. In order to stop full activity of inhibitory functions, it is almost always necessary to

inactivate both gene copies. The type of DNA change typical for tumour suppressor genes is DNA deletion. Often, one copy is lost by deletion while the second copy might be inactivated by base pair changes. Genetic alterations that predispose to tumour development are found in some hereditary cancer syndromes such as Li-Fraumeni syndrome (*TP53* mutations), Hereditary Breast and Ovarian Cancer Syndrome (*BRCA1/2* mutations), Cowden disease (*PTEN* mutations), and Von Hippel-Lindau disease (*VHL* mutations). In these syndromes, one copy of an inactivated tumour suppressor gene is present in the genome of all cells (germline) and therefore inherited, greatly increasing the chance of developing cancer.

The DNA changes described above are permanent alterations. However, also chemical base pair modification can occur and although reversible, can be quite stable and copied from parent to daughter cell (known as epigenetic alterations). An important epigenetic modification associated with cancer is DNA methylation. The frequent addition of a methyl group to cytosine in cytosine-guanine rich regions (referred to as CpG islands) located in promoter regions, will typically inhibit transcription of these genes (2, 3). DNA methylation is therefore an alternative mechanism to turn off tumour suppressor genes.

II. Hallmarks of cancer

Overcoming a series of cellular and bodily processes is thought to be necessary for the progression from a normal cell to a cancer cell. These may arise through genetic (mutation etc.) or epigenetic (e.g. DNA methylation, histone modification) means and include: (4)

- Self-sufficiency in growth signals
- Insensitivity to anti-growth signals
- Resistance to cell death
- Tissue invasion and metastases
- Limitless potential to replicate
- Inducing angiogenesis
- Evading the immune system

Late stage tumours have acquired all of these hallmarks. The methods used to acquire each hallmark will vary between cancers and in individuals, and reflects inherited or acquired pro-carcinogenic, events and individual features for each cancer. The pathway of acquiring these hallmarks is often called the Vogelstein model, after one of the first proponents (5) (Figure 1).

Almost all cells in our body are fully differentiated, and specialised to perform a limited number of tasks. These terminally differentiated cells have usually lost their potential to grow and so are unlikely to be the cells from which tumours develop. Tissue stem cells and precursor cells (to the differentiated mature cells) possess the capacity to divide

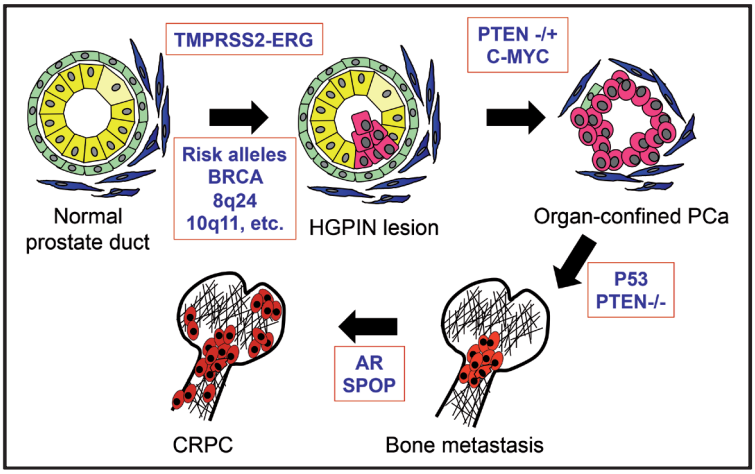


Figure 1: Evolution of prostate cancer. The development of high-grade prostatic intraepithelial neoplasia (HGPIN) and subsequent organ-confined prostate cancer is driven by genomic mutations. The presence of germline risk alleles such as *BRCA1/2* and gene polymorphisms increase the chance of developing prostate cancer. One of the first genomic events observed in the cancer cells is the fusion between *TMPRSS2* and *ERG* in about 50% of all tumours. Subsequent mutations include the loss of one *PTEN* allele and amplification of *c-MYC* on chromosome 8q24. In late stage tumours, the second *PTEN* allele is lost and *TP53* mutations are observed. After hormone therapy, androgen receptor (AR) amplification and mutations are observed in castration-resistant prostate cancer (CRPC). Recently, sequencing efforts identified *SPOP* as a mutated gene in late stage cancers.

and may be the origin of cancer. Except for the sporadic stem cells, all other cell types can only divide about 10-30 times when activated. During each DNA replication cycle, the chromosome ends (telomeres) are shortened. Only stem cells and cancer cells can repair this by expressing the enzyme telomerase.

In the absence of support factors, normal cells will stop dividing and become dormant (often by a process referred to as senescence) before they eventually die, in response to shortening telomeres. Many processes in the cell including the cell cycle, DNA damage response, protein expression, and energy metabolism include checkpoints to monitor potential harmful changes. Upon a challenge (e.g. DNA damage, too many incorrectly folded proteins), normal cells halt ongoing processes (stop dividing, stop transcription and translation, etc.), start repair mechanisms, and if repair is impossible, initiate a programmed cell death programme (apoptosis). Through random mutation and selection, cancer cells have typically mutated genes involved in these checkpoints to circumvent senescence and avoid apoptosis.

In prostate cancer (PCa), many common gene mutations are directly linked to adaptation of the hallmarks (Figure 1). The androgen receptor (AR) pathway becomes and remains the major growth stimulatory and anti-apoptotic cascade, partly by regulating the

TMPRSS2-ERG fusion gene (see below). Self-sufficient growth and resistance to cell death is further accomplished by loss of the tumour suppressor *PTEN* and amplification of the *c-MYC* oncogene. The tumour suppressor *TP53* can be altered to eliminate cell cycle arrest and prevent apoptosis (6).

III. Clinical aspects of prostate cancer

In Europe, PCa is the most common solid cancer among men. So far, age, ethnicity and heredity are the well-established risk factors for developing PCa (7). From the clinical point of view, one could subdivide interventions in three phases: the prevention phase, the detection and curative phase and the hormone, chemo and immunotherapy phase.

Regarding the prevention phase, several studies were perform to test whether nutritional supplementation as well as chemoprevention are valid prevention strategies for PCa (8). From the different trials performed so far, including the SELECT trial (selenium and vitamin E) and the PRP-1 trial (selenium, vitamin E and soy), the dietary intake of these supplements was either not recommendable or conferred no measurable improvement (9, 10). For chemoprevention, the most promising studies have tested the role of 5α-reductase inhibitors (finasteride and dutasteride) that block the conversion of testosterone (T) into 5α-dihydrotestosterone (DHT). Despite a reduction in the relative risk of PCa of about 25%, the occurrence of side-effects or even high-grade PCa in the group treated with these compounds revealed that population-based chemoprevention is not yet a feasible option (11, 12).

Of all PCa cases, only a minority is linked to inherited mutations in *BRCA2* (13) and although overall exogenous factors such as alcohol consumption and chronic inflammation might be involved in the development of PCa the evidence is not strong enough to recommend lifestyle changes (14).

In the early detection and curative phase, the main tools to diagnose PCa include the measurement of PSA (prostate specific antigen; KLK3) levels and digital rectal examination (DRE) (15). Higher levels of serum PSA are associated with an increased likelihood of PCa (16). Trials to use PSA levels to detect PCa in population-based screening programs failed in showing a clear benefit of screening for reducing PCa-related mortality that would outweigh side-effects including unnecessary biopsies and overtreatment (17). In the current diagnostic practice the need for a transrectal ultrasound (TRUS)-guided biopsy should not only be determined on the basis of the PSA level (7). PSA and DRE are neither sensitive enough to detect all relevant PCa cases nor specific enough to prevent unnecessary biopsies in case of prostatitis and benign prostatic hyperplasia (BPH) (18). There is still an urgent clinical need for more and better diagnostic PCa markers before biopsies are taken.

The diagnosis of PCa is based on histologic examination (19). From this point on, a decision is made regarding the primary local treatment of PCa. Patients with low-

risk and low-volume PCa can be advised for active surveillance. Patients with low- and intermediate- risk PCa and a life expectancy greater than 10 years have indication for radical prostatectomy (RP) or radiation therapy and low-dose-rate brachytherapy. Patients with the same risk but with a life expectancy less than 10 years can be advised towards watchful waiting. With respect to patients with high-risk clinically localized PCa there is no consensus on the optimal treatment but RP is often the first step. Alternatives include adjuvant external-beam radiation therapy (EBRT) and radiation therapy together with androgen deprivation therapy (ADT) (7).

In the hormone, chemo and immunotherapy phase, there is usually no curable endpoint. The objective is to slow any further progression of the disease or provide relief of disease symptoms through palliative care. The first-line treatment relies on androgen suppression by either bilateral orchiectomy or LHRH agonist/antagonist (20). When patients develop castration resistance, continuation of androgen suppression as well as further hormone and chemotherapies are the next step. In recent years, second- and third-line hormonal therapies have been successfully implemented including novel anti-androgens (such as enzalutamide) and CYP17 inhibitors (such as abiraterone acetate).

PCa, as a clinical entity, encompasses major needs in terms of better diagnostic and prognostic tools as well as novel treatment options. An adequate risk stratification of patients with newly diagnosed PCa will certainly provide a better support for treatment decisions. Also, the discovery of new relevant molecular targets for novel therapies can provide alternative treatment options to the individual patient.

IV. Major signalling pathways affected in prostate cancer

Major pathways targeted by mutations in PCa

In recent years the knowledge of molecular mechanisms of PCa growth and the role of individual genes and signalling pathways in this process, has rapidly increased. This knowledge is crucial for improvement of diagnosis and prognosis, and to offer patients more specific therapies. Like other cancers, PCa development and progressive growth is driven by an accumulation of variable numbers of genetic alterations in the cancer cell genome (Figure 1) (21). Moreover, it is becoming increasingly clear that epigenetic events, including the microenvironment of tumours, contribute substantially to regulation of tumour growth.

Approximately 5% of PCa cases have a familial history. Until recently, it was hardly possible to identify the hereditary alterations in the DNA, which are associated with tumour development. However, novel molecular technologies can now be instrumental to investigate this problem (see below). Genetically, sporadic PCa shows a characteristic pattern of chromosomal alterations. Most frequent changes are loss of the complete, or part of the short arm of chromosome 8 (8p), and losses of large regions of chromosome arms 6q, 13q and 16q. Additionally, in late stage PCa amplification of 8q is frequently

observed. In the affected chromosomal regions many candidate tumour suppressor genes and oncogenes are located. In most cases it is difficult to pinpoint exactly which of the genes or combination of genes are most important in tumour development, but it is well established that overexpression of c-MYC oncogene on 8q is involved in a large percentage of late stage prostate cancers.

Additional to large chromosomal regions, three small genomic regions on 10q, 21q and Xq, respectively, are frequently altered in PCa. Here, the genes important in PCa can easily be identified. These are the tumour suppressor gene *PTEN*, the *TMPRSS2-ERG* fusion gene and the *AR*.

In early stage PCa, one copy of *PTEN* is frequently lost; in late stage cancer both copies of the gene can be inactivated. *PTEN* is a phospholipid phosphatase that counteracts phosphoinositol-3-kinase (PI3K) (22). PI3K is activated by receptor tyrosine kinases, and following PI3K activation the phospholipid PIP2 can be modified to PIP3. *PTEN* catalyses the opposite conversion, PIP3 to PIP2. High concentrations of PIP3 in the cell lead to activation of the kinase AKT. Activation of this important PI3K/AKT signalling pathway affects many molecular and biological processes in the cell, including protein synthesis, cell proliferation, apoptosis, polarity and metabolism. The PI3K/AKT pathway can not only be activated by *PTEN* inactivation, but also by direct activation of PI3K or AKT. Remarkably, although occurring frequently in other tumour types, these alterations are rare in PCa (Figure 2).

The AR pathway in prostate cancer

DHT is essential for development, maintenance of the structure and function of the prostate. DHT is synthesized in the prostate by metabolism of testosterone by the enzyme 5-alpha-reductase. DHT mediates its function by activation of the AR, a member of the family of nuclear receptors, which are ligand-dependent transcription factors. Upon binding of DHT, the AR translocates from the cytoplasm to the nucleus. There it modulates gene expression by binding to specific DNA sequences, androgen response elements, in the promoter/enhancer regions of target genes (Figure 2).

In the normal prostate, AR expression is high in the luminal epithelial cells and absent in the basal epithelial cells. In stromal cells the level of AR expression is variable. It is generally accepted that AR in stromal cells is most important during prostate development, and AR in luminal cells for prostate function. Hundreds or even thousands of genes in the prostate are regulated by the activated AR. PSA, which is exclusively expressed in the luminal epithelial cells, is considered as the prototype of a tightly androgen-regulated gene. Castration has a dramatic effect on the normal prostate and results not only in modified gene expression, but also in enhanced apoptosis.

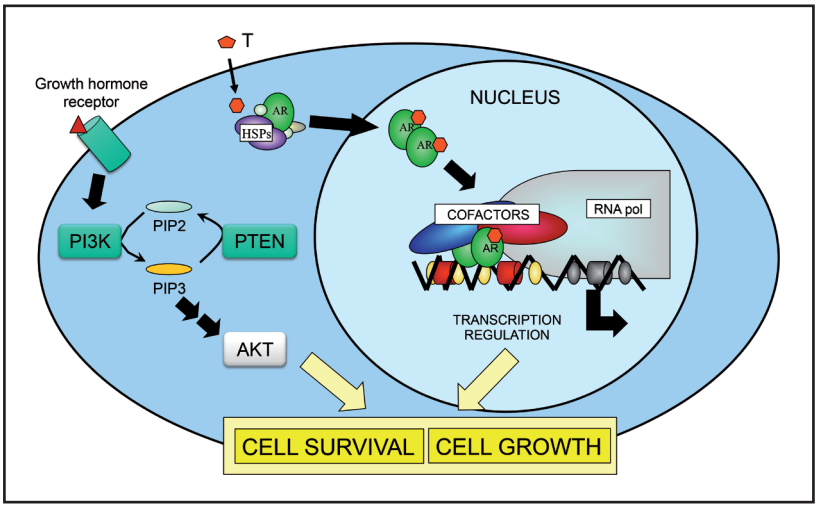


Figure 2: The *PTEN/PI3K/AKT* and AR pathways. PCa development and progression is driven by many different pathways of which two play a dominant role. The AR is activated by the androgens testosterone and dihydrotestosterone (DHT) after which the AR binds as a dimer to promoter and enhancer regions of target genes. One of the target genes is *TMPRSS2*, which is frequently fused to *ERG* in PCa, thereby regulating the *ERG* oncogene in an androgen-dependent manner. The PTEN/PI3K/AKT pathway is activated by various growth factors. At the inner cell membrane, activated PI3K eventually results in phosphorylation and activation of AKT (p-AKT). Loss of *PTEN* results in an increased and persistent activation of AKT. Both cascades support cell survival and cell growth and mutations (frequent *PTEN* loss, rare *PI3K* and *AKT* mutations, *AR* amplification and mutations, TMPRSS2-ERG fusion) affect these pathways frequently.

The TMPRSS2-ERG gene fusions in prostate cancer

The AR is vital for the development and function of the normal prostate and for growth of PCa. For a long time it was unknown which AR regulated genes were involved in tumour growth. In 2005, Tomlins *et al.* showed the presence of unique recurrent gene fusions between the *TMPRSS2* gene and *ETS* family members *ERG* and *ETV1* in PCa. Due to the gene fusion, the androgen-regulated *TMPRSS2* now controls the expression of the *ERG* or *ETV1* oncogene, having a large impact on androgen-dependent growth and survival of PCa (23).

TMPRSS2 and *ERG* are located on chromosome 21 approximately 3 Mb apart and have the same gene orientation. Mostly, *TMPRSS2-ERG* fusions are frequently accompanied by loss of the entire chromosome 21 sequence between *TMPRSS2* and *ERG* genes (interstitial deletion) (24). *TMPRSS2-ERG* has been found as the most common type of ETS gene fusion in PCa, with a frequency of approximately 50% across more than 1500 organ-confined PCa samples (25). In the PCa precursor lesion, high-grade prostatic intraepithelial neoplasia (HGPIN), the frequency was approx. 20% (26, 27). Additional

ETS gene fusions have subsequently been identified involving not only *ERG* and *ETV1* but also *ETV4* (28), *ETV5* (29), and *FLI1* (30). The frequency of these variants is much lower compared to ERG rearrangements but still accounts for 2-5% of all ETS gene fusions found in PCa.

In the normal situation, members of the ETS transcription factor family bind specific DNA sequences regulating the expression of nearby genes. They are believed to play a role in self-renewal-associated proliferation and normal morphogenesis (31, 32). The presence of ETS gene rearrangements in precursor lesions such as HGPIN suggests that ETS gene fusions have a causal role in the initial steps of PCa development (33). Still, it is not yet understood how ETS gene rearrangements are indeed capable of initiating prostate tumorigenesis whereas its role in mediating progression towards more aggressive disease is well accepted (34, 35).

Mismatch repair in prostate cancer

The DNA mismatch repair (MMR) system is a conserved mechanism that assists in maintaining genomic stability by correcting DNA mismatched base-pairs during the replication process (36). Defects in this repair mechanism are associated with mutations in microsatellite repeat sequences, which consist of tandemly repeated motifs of one (mono) to six (hexa) nucleotides (37) (38). The repetitive nature of these sequences makes them highly prone to insertion/deletion mutations caused by DNA polymerase slippage thereby altering the length of the microsatellite (39). The mismatch repair heterodimers MSH2/MSH6 and MSH2/MSH3 detect replication errors and recruit the MLH1/PMS2 complex which in turn degrades and resynthesizes the mutated stretch (40).

As a result, cells with a defective MMR system can have mutation rates 100-1000 fold higher than normal cells and frequently display microsatellite instability (MSI) (41). Several studies have shown that instability of microsatellites is present in PCa (42, 43). The frequency of microsatellite instability detected in PCa is variable between different studies and ranges from 8% to 35% (44, 45). Recently, an hypermutated phenotype has been observed in advanced prostate cancer patients which also displayed MSI and mutations in MMR genes (46). The expression of PMS2 also seems to be elevated in PCa patients and correlated with recurrence (47). It remains to be determined how frequently do rearrangements of MMR occur in PCa and whether defects in this pathway facilitate the transition of tumors to castration resistant disease.

Therapeutic targeting the AR pathway

More than 70 years ago, the beneficial effect of androgen ablation by castration on advanced PCa was described (48). Medical or surgical castration, aiming at inhibition of T and DHT production, has been the treatment of choice of metastatic PCa. Additional to castration, antiandrogens, e.g. flutamide and bicalutamide, which compete with DHT for AR binding, but do not activate AR, were added to the large panel of different endocrine treatments.

After a few years, essentially all tumours become refractory to endocrine therapy, and a resistant tumour emerges (castration-resistant prostate cancer (CRPC)). It turned out that CRPC still shows high expression of AR and AR target genes, including *TMPSS2-ERG*, indicating that AR is still important in this stage of the disease. However, the molecular mechanism of CRPC is complex, and might additionally to ERG, include a spectrum of many more genes for which the AR regulation has been altered (49).

During the last decades, a wide variety of mechanisms of escape from endocrine therapy have been described (50). It is now well established that in a proportion of CRPC (estimated 30%) *AR* on Xq is amplified, resulting in overexpression of AR mRNA and protein. Overexpressed *AR* can still be activated by low concentrations of androgens, or by androgens with low affinity for AR produced by the adrenal gland. In addition, there is evidence that overexpressed AR can function independent of an agonistic steroid ligand. Another way of AR activation in CRPC is by *AR* mutation. The best known mutation in *AR* is the threonine to alanine substitution in the AR ligand-binding domain, which allows the antagonist OH-Flutamide to display agonistic activity. More recently it was postulated that in CRPC constitutive active *AR* variants lacking the ligand-binding domain are present and can modify AR regulated gene expression. Moreover, evidence was provided that CRPC cells itself are able to synthesize T and DHT or are able to convert adrenal androgens to DHT.

During the last years, two novel endocrine therapies have been approved for CRPC treatment (51). One of the compounds recently marketed, enzalutamide or MDV3100, is related to more classical antiandrogens. Like other antiandrogens, it blocks AR function by competing with DHT for AR binding. However, enzalutamide also delays nuclear transport of AR. Although enzalutamide has been approved for treatment of CRPC patients following chemotherapy with docetaxel, at the moment there is no reason to believe that it is not effective in earlier stages of PCa. The second novel therapeutic compound is abiraterone acetate. This molecule inhibits the function of CYP17A1, a key enzyme of the DHT synthesis pathway. Abiraterone acetate not only inhibits T synthesis in the testis but also the synthesis of adrenal androgens. Abiraterone acetate is beneficial in late stage PCa patients resistant to docetaxel therapy. Abiraterone acetate is given in combination with prednisone to reduce its side effects, which seem more severe than for enzalutamide.

V. Technology developments and emerging markers

Increasing knowledge of the molecular biology of PCa and breakthroughs in various technologies, form the basis of the discovery of novel markers and therapy targets. The technologies advanced tremendously in the ability to measure more of the DNA, RNA and protein content of each individual tumour (high-content) and allow analysis of more samples at the same time (high-throughput). These type of studies are referred to as '-omics'; for DNA, RNA and protein analyses, genomics, transcriptomics and proteomics, respectively.

Genomics and Transcriptomics

Miniaturization of assays has been the main driving force for the high-content detection of RNA and DNA using array technologies and next generation sequencing (NGS). Microarrays can be used as a read-out system for gene expression by hybridizing processed RNAs or for changes in DNA copy number by hybridizing genomic DNA, referred to as array CGH (comparative genomic hybridization) (52). Only recently, massive parallel sequencing techniques have become affordable and a more comprehensive alternative to array-based analyses. Sequencing of RNA not only identifies changes in gene expression, but also mutations, alternative splicing and gene fusion events. At the DNA level, sequencing of all 3.2 billion base pairs in human samples at high coverage is still challenging with respect to costs, data handling and processing (53). However, since the sequencing technologies are progressing rapidly, it is expected that within the next decade, whole genome sequencing can be performed in a few days for a thousand pounds. If so, the implementation of NGS in the clinic for mutation analysis of PCa will likely become reality. Using NGS, individual prostate samples can be screened for diagnostic and prognostic markers and therapy targets to determine on a personal basis, the significance of the disease, the treatments to which the patient will respond and the optimal order of therapies (54, 55).

Proteomics

Large scale research on proteins is much more complex than genomics and transcriptomics. Instead of the four bases, proteins are build-up by 20 different amino acids that can also be modified in many different ways (e.g. phosphorylation, glycosylation). Proteins cannot be amplified like RNA and DNA using the polymerase chain reaction (PCR) and protein detection technologies, mainly based on antibodies, often lack sensitivity and specificity. The breakthrough in protein detection came with the advancements of mass spectrometry (MS). This technology can accurately measure the mass of proteins and protein fragments and also allows for determining the amino acid sequence of a selected protein fragment (56). With some effort, hundreds to thousands of different proteins can be identified in tissue extracts or body fluids such as serum or urine. However, low abundant proteins in complex mixtures are typically difficult to detect using MS and in general, blood-based cancer markers such as PSA are present in low quantities. The implementation of MS technologies for marker research has not resulted in a wide range of novel clinically useful protein markers. Reasons are the limitations of MS to detect low abundant proteins and the huge variability in protein composition of body fluids between individuals. MS technology is nowadays slowly improving the detection of low abundant proteins and high-throughput measurements and eventually, MS is expected to become a technique operational in a clinical setting for the detection of protein biomarkers.

Current and new markers and targets based on our knowledge of prostate cancer biology

For the diagnosis of PCa, a limited number of markers are commonly in use. These include family history, serum PSA, DRE and imaging technologies (e.g. MRI) that provide an indication of prostate abnormalities including cancer. The final diagnosis is established by histopathological examination of prostate needle biopsies. The biopsies provide information for prognosis which includes an estimate of the extent of the cancer (number of positive cores and % cancer area in each core) and particularly the differentiation state of the cancer cells, which is graded using the Gleason score. Both the diagnosis and prognosis of PCa are far from perfect and there is a strong clinical need for additional markers.

Since ETS gene fusions are the most frequent genetic abnormalities found in PCa, there is considerable interest in using ETS fusion transcripts as diagnostic and prognostic markers [18]. ETS fusion events are unique to cancer and their detection has high diagnostic value. Molecular tests to measure *TMPRSS2-ERG* transcripts in urine after DRE have been developed and are in the final stages of clinical testing. With respect to their prognostic value, conflicting studies report that ERG fusions are associated with both favourable and unfavourable clinical outcomes [21]. It is unlikely that ERG overexpression by itself will be a useful prognostic marker.

Diagnostic RNA markers

Besides *TMPRSS2-ERG*, other transcript markers are making their way into the clinic. PCA3, a long noncoding RNA was discovered in 1999 as a prostate-specific molecule, upregulated in PCa (57). A commercial test has been developed (Progensa) to measure PCA3 and PSA mRNA levels in whole urine after DRE. The PCA3 score is calculated as the ratio of PCA3/PSA mRNA x 1000. The diagnostic role of PCA3 has been studied in the setting of predicting biopsy outcome, mainly to prevent unnecessary repeat biopsies of men with previous negative biopsies but persisting elevated PSA levels. It was shown that PCA3 scored better than PSA alone and complements established PCa risk factors such as age, PSA, DRE and prostate volume, making PCA3 a strong candidate to add to existing risk calculators (58).

As a prognostic marker, PCA3 has been limited in its ability to predict PCa stage and aggressiveness beyond established markers (58). Since PCa is a very heterogeneous and often multifocal disease, combining different cancer markers, such as *TMPRSS2-ERG* and PCA3, is expected to improve diagnostic accuracy. Overall, combining new urine biomarkers that are in the pipeline with the PCA3 score might significantly improve predictability of biopsy outcome (58).

Predictive markers

In PCa, little progress has been made in developing predictive markers to aid in therapy decision. With the AR as the major target regarding PCa treatment, recent studies have shown promising results regarding the presence of *AR* variants and response to treatment (59). Resistance to abiraterone and enzalutamide has been associated with an increase in expression of certain *AR* variants that render the receptor constitutively active (60). With the development of compounds that target the N-terminal portion of the AR, such as EPI-001, detection of *AR* variants can be coupled with a change in the treatment regimen (59, 61).

Targeting DNA repair mechanisms in cancer treatment has been widely used in the clinic for years as a way to circumvent acquired chemo- and radio-resistance (62). In PCa, PARP-1 inhibitors were shown to cooperate with castration therapy enhancing the therapeutic response (63). Also *in vitro*, the use of PARP-1 inhibitors was capable of suppressing CRPC cell growth. Further research is required to confirm the therapeutic improvement of using PARP-1 inhibitors in PCa patients with *BRCA1/2* mutations.

Prognostic RNA markers

Many initiatives are being undertaken to identify and validate prognostic RNA markers. Based on knowledge of changes in cell cycle genes in cancer, a cell cycle progression (CCP) score, created from expression levels of 31 transcripts was tested and validated (64). This profile requires RNA extracted from cancer tissue and should be performed on positive biopsies. Expression profiling of RNA isolated from whole blood has also been shown to separate men with CRPC in groups with long versus short median survival (65, 66). These whole blood profiles most likely reflect changes in blood cells and not in the rare circulating tumour cells (CTCs) or tumour-derived RNAs. Whether the prognostic power of any of these mRNA profiles persist in large independent validation cohorts still needs to be determined.

Another type of RNA that received much attention for its regulatory role in cancer are short (~22 nt) RNAs referred to as microRNAs (miRNAs). These miRNAs are rather stable and resistant to sample treatments (e.g. fixation, long term storage) and therefore attractive candidate markers. Several recent studies have successfully indicated miRNAs as diagnostic and prognostic markers in tissue and blood (67, 68). Like for mRNAs mentioned above, a tissue profile of 4 miRNAs ((miR-96 x miR-183)/(miR-145 x miR-221)) was shown to improve accuracy and outperform PSA-based diagnosis and predict PCa aggressiveness (69).

Emerging markers from –omics studies

Recent large scale NGS projects provided an overview of mutations in PCa (70-72). As expected, most frequently detected mutated genes were known as part of the AR pathway, PTEN/AKT axis or cell cycle regulatory genes. However, tens of novel

mutated genes were identified that have a very low occurrence rate (1-3%) in PCa and a few new mutated genes that are more commonly affected (4-15%). One previously unknown common mutated gene is SPOP, encoding a protein that is involved in protein degradation and may modulate transcriptional repression. For many of these newly discovered mutated genes, the research on their role as markers and therapy targets has now been initiated.

Gene fusions/mutations

The emerging number of next generation sequencing studies extended our knowledge of the genomic landscape in PCa. In 2011, the first whole genome sequencing report in PCa identified several rearrangements involving the *PTEN* gene and mutations in genes involved in chromatin modifications and heat shock protein pathways (73). Exome sequencing identified novel mutations associated with PCa, in particular, *SPOP* was found mutated in 6-15% of tumours that lacked ETS gene rearrangements (71). Additional exome sequencing studies focused in CRPC disease detected frequent copy number losses in CHD1 that defined a novel subtype of ETS negative PCa (72). Also mutations have been detected to a certain extend in *FOXA1*, a known AR co-regulator (71, 72). The nuclear receptor coactivator *NCOA2* was found amplified in ~11% of tumors (74). Several novel gene fusions have been detected in PCa, but mostly these are individual to each patient. Gene rearrangements involving genes of the RAF pathway were detectable in 1–2% of prostate cancers (75).

Together, the recent next generation sequencing studies have highlighted heterogeneity as the most dominant feature in PCa. Common genetic abnormalities like ETS gene fusions and *PTEN* loss co-occur with many other alterations that appear to be patient specific.

Exosomes

As described above, on protein level, candidate markers have been discovered in PCa tissue, urine and serum using mass spectrometry (76). Thorough independent validation of almost all of these candidate markers still needs to be performed. One of the interesting findings from MS experiments on proteins secreted by cancer cells, is the finding that many of these proteins are secreted via small vesicles, often referred to as exosomes (77, 78). This exosomal secretory pathway is completely different from the well-known endoplasmic reticulum-Golgi route that proteins such as PSA follow to be released from cells (Figure 3).

Exosomes are small (50-100 nm) vesicles which are formed by inward budding of cytoplasmic endosomes (79). An endosome containing exosomes is called a multivesicular body (MVB) and when the MVB fuses with the cellular membrane the exosomes are secreted. Exosomes are secreted by most cell types and also the normal prostate produces large quantities of exosome-like vesicles called prostasomes (80). Exosomes contain the proteins and RNAs from the cytoplasm of the cell of origin (81, 82). This provides the unique opportunity to study the presence and molecular

properties of cancer cells by analyzing the protein and RNA content of cancer-derived exosomes. Like for PSA, the presence of prostasomes in blood is a marker for PCa (83). Also in urine, extracellular vesicles from the prostate can be measured and the signal of the exosomal markers CD9 and CD63 are significantly higher in men with PCa (84).

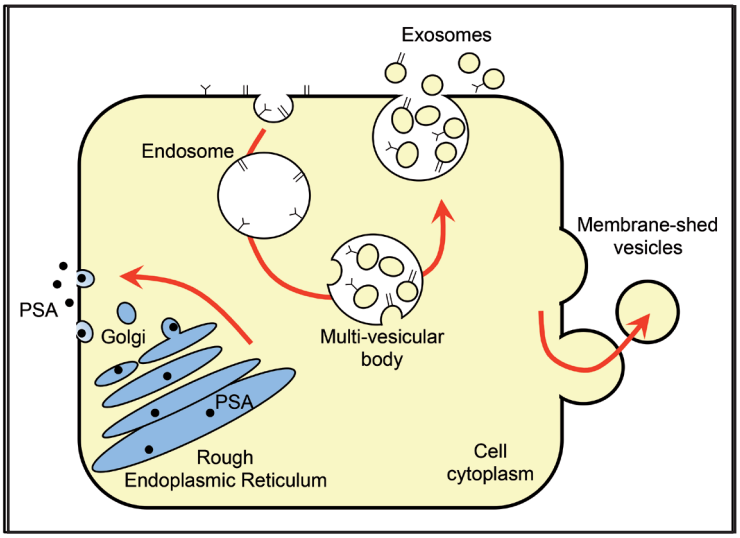


Figure 3: Overview of different secretion mechanisms. Proteins with an N-terminal signal peptide are translated at the rough endoplasmic reticulum and transported via the Golgi apparatus to the cell membrane for secretion into the milieu.

Genome-wide association studies and genetic variation

It is known for a long time that approx. 5% of prostate cancers are clearly linked to a family history. Typically, genome-wide association studies (GWAS) are used to identify the presence of common variants (single nucleotide polymorphisms - SNPs) in chromosomal regions and genes that influence complex diseases. In PCa, 78 published loci (Table 1) have been associated with prostate cancer ($P < 5 \times 10^{-8}$) but this only explains 31% of the PCa heritability (85-86). The number of genes identified by GWAS that are likely causally linked to hereditary PCa is limited. Besides these common variants, rare germline variants have been identified within families, explaining another part of hereditary PCa. Mutations in the BRCA2 gene that is very important in hereditary breast cancer, are associated with a very small percentage of hereditary PCa (87). Recently however, it has been discovered that a specific mutation in *HOXB13* (HOXB13-G84E) is associated with approx. 5% of hereditary PCa (88). The mutation is also present in less than 1% of PCa patients without a known family history. The mutation is approx. 10-fold more common in PCa patients than in normal controls. Importantly, now that rapid and affordable whole genome sequencing is possible, more common genomic polymorphisms such as SNPs, deletions and insertions will be identified and possibly

linked to hereditary PCa. Hopefully, this will lead to the identification of novel genes that at low- or high-penetrance, predispose to the development of high risk PCa.

Table 1: List of the published loci associated with prostate cancer risk

SNP	Locus	Candidate gene	Risk allele frequency	Odds ratio	Reference
rs1218582	1q21	*KCNN3*	0.45	1.06	Eeles *et al*, 2013
rs4245739	1q32	*MDM4–PIK3C2B*	0.25	0.91	Eeles *et al*, 2013
rs10187424	2p11	*GGCX—VAMP8*	0.41	0.92	Kote-Jarai *et al*, 2011
rs721048	2p15	*EHBP1*	0.19	1.15	Gudmundsson *et al*, 2010
rs6545977	2p15	*OTX1—RPL27P5*	NR	NR	Eeles *et al*, 2013
rs1465618	2p21	*THADA*	0.23	1.08	Eeles *et al,* 2013
rs13385191	2p24	*C2orf43*	0.4	1.15	Takata *et al*, 2010
rs11902236	2p24	*TAF1B:GRHL1*	0.27	1.07	Eeles *et al*, 2013
rs12621278	2q31	*ITGA6*	0.06	0.75	Eeles *et al*, 2013
rs3771570	2q37	*FARP2*	0.15	1.12	Eeles *et al*, 2013
rs7584330	2q37	*COL6A3 - MLPH*	0.22	1.06	Kote-Jarai *et al*, 2011
rs7629490	3p11	*VGLL3 - CHMP2B*	NR	1.06	Schumacher *et al*, 2011
rs9284813	3p12	*VGLL3 - CHMP2B*	NR	NR	Takata *et al*, 2010
rs2660753	3p12		0.11	1.18	Kote-Jarai *et al*, 2011
rs7611694	3q13	*SIDT1*	0.41	0.91	Eeles *et al*, 2013
rs10934853	3q21	*GATA2*	0.4513	1.12	Gudmundsson *et al*, 2010
rs4857841	3q21	*EEFSEC*	0.3	1.13	Lindstrom *et al*, 2012
rs9311171	3p22	*CTDSPL*	0.1997	NR	Murabito *et al*, 2007
rs6763931	3q23	*ZBTB38*	0.45	1.04	Kote-Jarai *et al*, 2011
rs345013	3q24	*PLOD2*	0.1102	NR	Murabito *et al*, 2007
rs1894292	4q13	*AFM,RASSF6*	0.48	0.91	Eeles *et al*, 2013
rs17021918	4q22	*PDLIM5*	0.34	0.9	Eeles *et al*, 2013
rs12500426	4q22	*PDLIM5*	0.46	1.08	Eeles *et al*, 2013
rs7679673	4q24	*RPL6P14–TET2*	0.45	0.91	Eeles *et al*, 2013
rs2121875	5p12	*FGF10*	0.34	1.05	Kote-Jarai *et al*, 2011
rs2242652	5p15	*TERT*	0.19	0.87	Kote-Jarai *et al*, 2011
rs12653946	5p15	*IRX4 - IRX2*	0.5	1.31	Takata *et al*, 2010
rs4466137	5q14	*HAPLN1*	NR	NR	Murabito *et al*, 2007
rs2242652	5q15	*LPCAT1*	0.1768	1.15	Kote-Jarai *et al*, 2011
rs6869841	5q35	*FAM44B (BOD1)*	0.21	1.07	Eeles *et al*, 2013
rs2273669	6p21	*ARMC2,SESN1*	0.15	1.07	Eeles *et al*, 2013
rs1983891	6p21	*FOXP4*	0.41	1.15	Takata *et al*, 2010
rs130067	6p21	*CCHCR1*	0.21	1.05	Kote-Jarai *et al*, 2011
rs339331	6q22	*GPRC6A;RFX6*	0.31	1.28	Takata *et al*, 2010
rs1933488	6q25	*RSG17*	0.41	0.89	Eeles *et al*, 2013
rs651164	6q25	*LOC100289162*	NR	0.87	Schumacher *et al*, 2011
rs9364554	6q25	*SLC22A3*	0.29	1.17	Eeles *et al*, 2013
rs12155172	7p15	*RPS26P30–ASS1P11-SP8*	0.2	1.05	Eeles *et al*, 2013
rs10486567	7p15	*JAZF1*	0.77	0.74	Thomas *et al*, 2008
rs6465657	7q21	*LMTK2*	0.46	1.12	Eeles *et al*, 2013

Table 1: Continued

rs2928679	8p21	*SLC25A37 NKX3-1*	0.42	1.05	Eeles *et al*, 2013
rs1512268	8p21	*SLC25A37 - NKX3-1*	0.45	1.18	Eeles *et al*, 2013
rs11135910	8p21	*EBF2*	0.16	1.11	Eeles *et al*, 2013
rs10086908	8q24	*POU5F1B, MYC*	0.3	0.87	Eeles *et al*, 2013
rs12543663	8q24	*LOC727677, MYC*	0.33	1.08	Al Olama AA *et al*, 2009
rs13252298	8q24	*FAM84B - SRRM1P1*	NR	0.89	Schumacher *et al*, 2011
rs445114	8q24	*SRRM1P1 - POU5F1B*	0.36	1.14	Gudmundsson *et al*, 2010
rs16902094	8q24	*SRRM1P1 - POU5F1B*	0.15	1.21	Gudmundsson *et al*, 2010
rs817826	9q31	*RAD23B-KLF4*	0.1	1.43	Xu *et al*, 2010
rs1571801	9q33	*DAB2IP*	0.25	1.27	Duggan *et al*, 2007
rs10933994	10q11	*MSMB - NCOA4*	0.4	1.25	19 or 30
rs2252004	10q26	*NR*	0.23	1.16	Akamatsu *et al*, 2012
rs11199874	10q26	*RPL19P16-FGFR2*	0.29	2.9	Nam *et al*, 2006
rs4962416	10q26	*CTBP2*	0.27	1.2	Thomas *et al*, 2008
rs7127900	11p15	*IGF2-INS*	0.2	1.22	Eeles *et al*, 2013
rs1938781	11q12	*FAM111A*	0.3	1.16	Akamatsu *et al*, 2012
rs11228565	11q13	*TPCN2 - MYEOV*	0.2	1.23	Gudmundsson *et al*, 2010
rs11568818	11q22	*MMP7*	0.44	0.91	Eeles *et al*, 2013
rs731236	12q13	*VDR*	0.264	NR	Bonilla *et al*,2011
rs10875943	12q13	*TUBA1C-PRPH*	0.31	1.07	Kote-Jarai *et al*, 2011
rs12827748	12q21	*PAWR*	0.312	NR	Bonilla *et al*, 2011
rs1270884	12q24	*TBX5*	0.49	1.07	Eeles *et al*, 2013
rs1529276	13q33	*SLC10A2-RPL7P45*	NR	NR	Murabito *et al*, 2007
rs8008270	14q22	*FERMT2*	0.18	0.89	Eeles *et al*, 2013
rs7141529	14q24	*RAD51B*	0.5	1.09	Eeles *et al*, 2013
rs4775302	15q21	*SQRDL*	0.4293	1.41	Nam *et al*, 2011
rs684232	17q13	*VPS53*	0.4881	1.1	Eeles et al, 2013
rs11650494	17q21	*GNGT2, ABI3, PHB, SPOP, HOXB13*	0.08	1.15	Eeles *et al*, 2013
rs7241993	18q23	*SALL3*	0.3	0.92	Eeles *et al*, 2013
rs8102476	19q13	*DPF1 - PPP1R14A*	0.46	1.12	Gudmundsson *et al*, 2010
rs103294	19q13	*LILRA3*	0.3	1.28	Xu *et al*, 2010
rs2427345	20q13	*GATAS, CABLES2*	0.37	0.94	Eeles *et al*, 2013
rs6062509	20q13	*ZGPAT*	0.3	0.89	Eeles *et al*, 2013
rs9623117	22q13	*TNRC6B*	0.3058	1.18	Sun *et al*, 2009
rs742134	22q13	*BIK*	NR	1.16	Schumacher *et al*, 2011
rs5759167	22q13	*RPS25P10-BIK*	0.47	0.86	Eeles *et al*, 2013
rs5945619	Xp11	*NUDT11*	0.36	1.19	Eeles *et al*, 2013
rs5919432	Xq12	*AR*	0.19	0.94	Eeles *et al*, 2013

VI. Scope of the thesis

Prostate cancer is one of the most common cancers among men. Despite relatively low morbidity, this disease represents a major burden in the current western society. In PCa, gene fusions are among the most frequently observed genetic alterations. Therefore, the aim of this thesis was to identify novel DNA and RNA gene fusions in PCa that, just as *TMPRSS2-ERG* could provide novel insight into the initiation and/or progression pathways of PCa.

In chapter 2 we combined DNA and RNA array data of several PCa samples as a first attempt to detect frequent unique gene fusions in PCa. The rationale was to select the intragenic deletions and amplifications that coincided with altered expression patterns. We were able to identify a novel gene fusion between the genes *GPS2* and *MPP2* that resulted in an in-frame chimeric protein. If this gene fusion conferred a selective advantage regarding PCa progression, we expected *GPS2-MPP2* to contribute to the proliferation or apoptosis of the cancer cells, which we later confirmed. Since we were not able to detect this particular gene fusion in any of the other PCa samples and the validation process of gene fusions using the DNA and RNA arrays was complex, we next utilized DNA and RNA sequencing data for fusion detection. In chapter 3, a catalogue all the gene fusions present in the PCa cell line PC346C and the PCa patient sample G089 was generated using whole genome sequencing data. We detected another gene fusion involving a membrane palmitoylated protein, in this case, *MPP5. MPP5* was fused to *FAM71D*, which was overexpressed in the cancer cells harbouring the gene fusion. Also *MPP5-FAM71D* showed to be relevant for the proliferation capacity of the cells and despite the uniqueness of this gene fusion we did find overexpression of *FAM71D* in a subset of PCa samples.

Mismatch repair (MMR) deficiency has been described in PCa but the exact role the MMR system plays in PCa, is not well established. In order to determine whether the deficiency in the MMR system might cause advantageous gene mutations in the context of PCa we analysed, in chapter 4, the genome of the PCa MMR-deficient cell line PC346C. Although mutations in the MMR system are so far relatively rare in PCa, they might represent a different subgroup of cancers and therefore be of relevance in patient stratification. We predicted that nucleotide repeat sequences would be more frequently hit by mutations in a MMR-deficient background. With this approach we found 14 candidate genes, of which *PRRT2* was frequently mutated in microsatellite unstable colorectal and endometrial cancer patients, but not in PCa. Although the *PRRT2* mutation in the context of PCa has so far only been observed in the PC346C cell line, knock-down of the truncated *PRRT2* protein did decrease proliferation of these cells.

In light of the new discoveries regarding the mechanisms that generate genetic diversity, in particular chromothripsis, we hypothesized in chapter 5 that such a mechanism could also play a role in PCa. After analysis of all the structural rearrangements present in the sequenced PCa samples, we found VCaP to have chromothripsis in the chromosome 5q arm. Focusing once again on gene fusions we questioned whether the occurrence of

chromothripsis would originate in a high number of functional gene fusion events and therefore be a selective advantage to the cancer cell. Compared to the total number of rearrangements present in the chromotripic arm, only a minor fraction resulted in in-frame gene fusions. We concluded that chromothripsis is unlikely responsible for the generation of relevant gene fusions in PCa.

In an attempt to broaden our knowledge of PCa biology we have discovered novel genetic alterations in genes and pathways not previously linked to PCa. We have also shown that chromothripsis occurs in PCa. In chapter 6 we discuss the relevance of the new findings in the context of our basic understanding and clinical aspect of this disease.

References

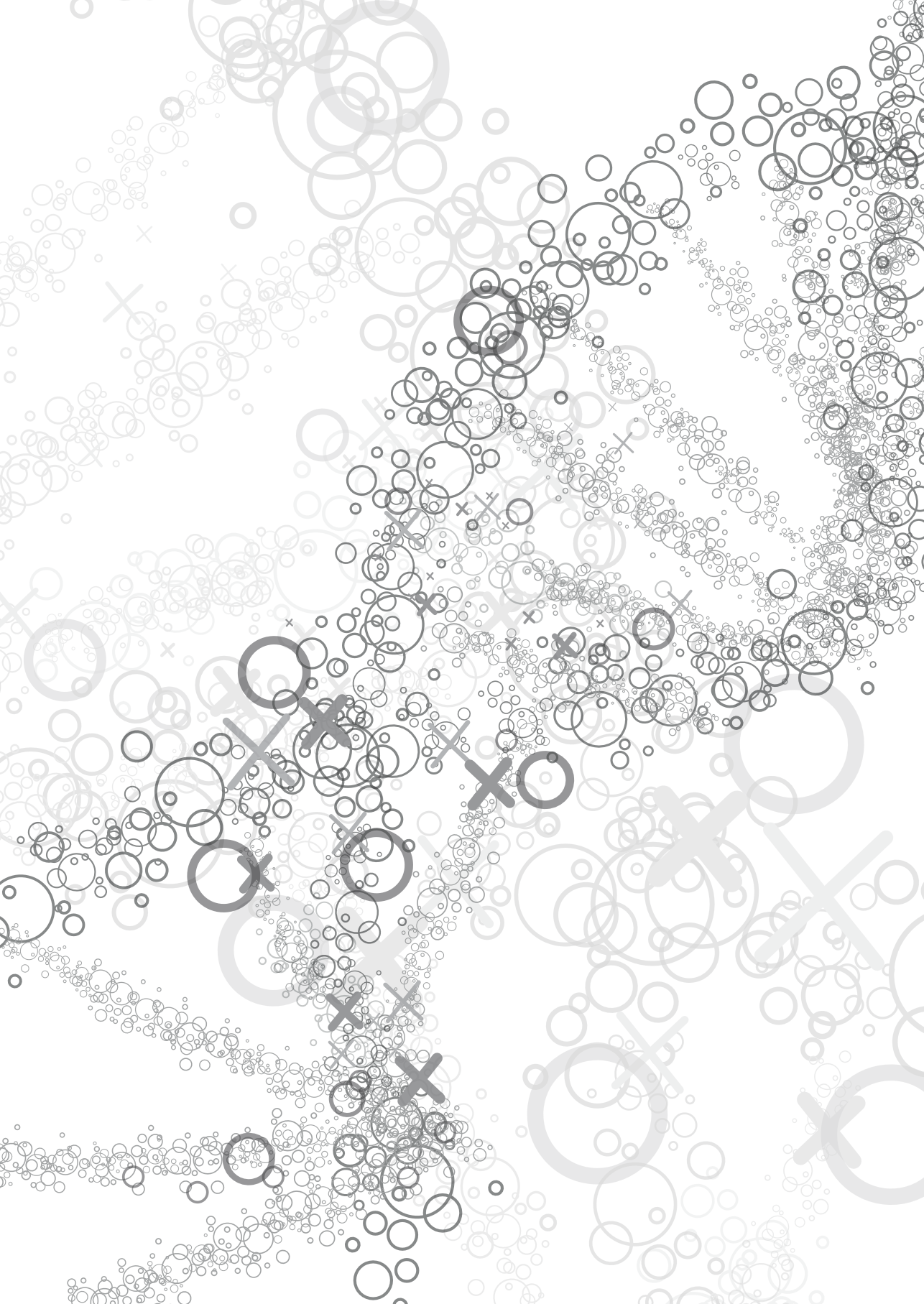
1. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-74.
2. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683-92.
3. Sandoval J, Esteller M. Cancer epigenomics: beyond genomics. Curr Opin Genet Dev. 2012;22(1):50-5.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
5. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5): 759-67.
6. Oren M, Rotter V. Introduction: p53--the first twenty years. Cell Mol Life Sci. 1999;55(1):9-11.
7. Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, van der Kwast T, et al. EAU guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent-update 2013. Eur Urol. 2014;65(1):124-37.
8. Pinto A, Perez Segura P. Primary prevention and early diagnosis of prostate cancer: recommendations from the prevention and early diagnosis working group of the Spanish Society of Medical Oncology (SEOM). European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation. 2015.
9. Klein EA, Thompson IM, Jr., Tangen CM, Crowley JJ, Lucia MS, Goodman PJ, et al. Vitamin E and the risk of prostate cancer: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). Jama. 2011;306(14):1549-56.
10. Fleshner NE, Kapusta L, Donnelly B, Tanguay S, Chin J, Hersey K, et al. Progression from high-grade prostatic intraepithelial neoplasia to cancer: a randomized trial of combination vitamin-E, soy, and selenium. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2011;29(17):2386-90.
11. Thompson IM, Goodman PJ, Tangen CM, Lucia MS, Miller GJ, Ford LG, et al. The influence of finasteride on the development of prostate cancer. The New England journal of medicine. 2003;349(3):215-24.
12. Andriole GL, Bostwick DG, Brawley OW, Gomella LG, Marberger M, Montorsi F, et al. Effect of dutasteride on the risk of prostate cancer. The New England journal of medicine. 2010;362(13):1192-202.
13. Bratt O. Hereditary prostate cancer: clinical aspects. The Journal of urology. 2002;168(3): 906-13.
14. Leitzmann MF, Rohrmann S. Risk factors for the onset of prostatic cancer: age, location, and behavioral correlates. Clin Epidemiol. 2012;4:1-11.
15. Lim LS, Sherin K, Committee APP. Screening for prostate cancer in U.S. men ACPM position statement on preventive practice. Am J Prev Med. 2008;34(2):164-70.
16. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. The New England journal of medicine. 2004;350(22):2239-46.
17. Cuzick J, Thorat MA, Andriole G, Brawley OW, Brown PH, Culig Z, et al. Prevention and early detection of prostate cancer. Lancet Oncol. 2014;15(11):e484-92.
18. Prensner JR, Rubin MA, Wei JT, Chinnaiyan AM. Beyond PSA: the next generation of prostate cancer biomarkers. Sci Transl Med. 2012;4(127):127rv3.

19. van der Kwast TH, Lopes C, Santonja C, Pihl CG, Neetens I, Martikainen P, et al. Guidelines for processing and reporting of prostatic needle biopsies. J Clin Pathol. 2003;56(5):336-40.
20. Horwich A, Parker C, de Reijke T, Kataja V, Group EGW. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2013;24 Suppl 6:vi106-14.
21. Shen MM, Abate-Shen C. Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev. 2010;24(18):1967-2000.
22. Song MS, Salmena L, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor. Nat Rev Mol Cell Biol. 2012;13(5):283-96.
23. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310(5748):644-8.
24. Hermans KG, van Marion R, van Dekken H, Jenster G, van Weerden WM, Trapman J. TMPRSS2:ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer. Cancer Res. 2006;66(22):10658-63.
25. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008;8(7):497-511.
26. Cerveira N, Ribeiro FR, Peixoto A, Costa V, Henrique R, Jeronimo C, et al. TMPRSS2-ERG gene fusion causing ERG overexpression precedes chromosome copy number changes in prostate carcinomas and paired HGPIN lesions. Neoplasia. 2006;8(10):826-32.
27. Perner S, Mosquera J-M, Demichelis F, Hofer MD, Paris PL, Simko J, et al. TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion. Am J Surg Pathol. 2007;31(6):882-8.
28. Tomlins SA, Mehra R, Rhodes DR, Smith LR, Roulston D, Helgeson BE, et al. TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. Cancer Res. 2006;66(7):3396-400.
29. Helgeson BE, Tomlins SA, Shah N, Laxman B, Cao Q, Prensner JR, et al. Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. Cancer Res. 2008;68(1): 73-80.
30. Paulo P, Barros-Silva JD, Ribeiro FR, Ramalho-Carvalho J, Jeronimo C, Henrique R, et al. FLI1 is a novel ETS transcription factor involved in gene fusions in prostate cancer. Genes Chromosomes Cancer. 2012;51(3):240-9.
31. Sashida G, Bazzoli E, Menendez S, Liu Y, Nimer SD. The oncogenic role of the ETS transcription factors MEF and ERG. Cell Cycle. 2010;9(17):3457-9.
32. Rosen P, Sesterhenn IA, Brassell SA, McLeod DG, Srivastava S, Dobi A. Clinical potential of the ERG oncoprotein in prostate cancer. Nat Rev Urol. 2012;9(3):131-7.
33. Clark J, Attard G, Jhavar S, Flohr P, Reid A, De-Bono J, et al. Complex patterns of ETS gene alteration arise during cancer development in the human prostate. Oncogene. 2008;27(14):1993-2003.
34. Carver BS, Tran J, Chen Z, Carracedo-Perez A, Alimonti A, Nardella C, et al. ETS rearrangements and prostate cancer initiation. Nature. 2009;457(7231):2-3.
35. Tomlins SA, Bjartell A, Chinnaiyan AM, Jenster G, Nam RK, Rubin MA, et al. ETS gene fusions in prostate cancer: from discovery to daily clinical practice. Eur Urol. 2009;56(2):275-86.
36. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008;18(1):85-98.

37. Shah SN, Hile SE, Eckert KA. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res. 2010;70(2):431-5.
38. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. Trends in biotechnology. 2007;25(11):490-8.
39. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature reviews Genetics. 2004;5(6):435-45.
40. Pena-Diaz J, Jiricny J. Mammalian mismatch repair: error-free or error-prone? Trends in biochemical sciences. 2012;37(5):206-14.
41. Peltomaki P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Human molecular genetics. 2001;10(7):735-40.
42. Kagan J, Pisters L, Troncoso P, Joe Y, Parat J, Babaian R, et al. Genetic instability in microsatellite sequences in prostate-cancer. International journal of oncology. 1994;5(4):921-4.
43. Watanabe M, Imai H, Shiraishi T, Shimazaki J, Kotake T, Yatani R. Microsatellite instability in human prostate cancer. British journal of cancer. 1995;72(3):562-4.
44. Dahiya R, Lee C, McCarville J, Hu W, Kaur G, Deng G. High frequency of genetic instability of microsatellites in human prostatic adenocarcinoma. International journal of cancer Journal international du cancer. 1997;72(5):762-7.
45. Perinchery G, Nojima D, Goharderakhshan R, Tanaka Y, Alonzo J, Dahiya R. Microsatellite instability of dinucleotide tandem repeat sequences is higher than trinucleotide, tetranucleotide and pentanucleotide repeat sequences in prostate cancer. International journal of oncology. 2000;16(6):1203-9.
46. Pritchard CC, Morrissey C, Kumar A, Zhang X, Smith C, Coleman I, et al. Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. Nat Commun. 2014;5.
47. Norris AM, Gentry M, Peehl DM, D'Agostino R, Scarpinato KD. The elevated expression of a mismatch repair protein is a predictor for biochemical recurrence after radical prostatectomy. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2009;18(1):57-64.
48. Huggins C. EFFECT OF ORCHIECTOMY AND IRRADIATION ON CANCER OF THE PROSTATE. Ann Surg. 1942;115(6):1192-200.
49. Sharma NL, Massie CE, Ramos-Montoya A, Zecchini V, Scott HE, Lamb AD, et al. The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. Cancer Cell. 2013;23(1):35-47.
50. Lamont KR, Tindall DJ. Minireview: Alternative activation pathways for the androgen receptor in prostate cancer. Mol Endocrinol. 2011;25(6):897-907.
51. Attard G, Richards J, de Bono JS. New strategies in metastatic prostate cancer: targeting the androgen receptor signaling pathway. Clin Cancer Res. 2011;17(7):1649-57.
52. Dufva M. Introduction to microarray technology. Methods Mol Biol. 2009;529:1-22.
53. Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. Annu Rev Genomics Hum Genet. 2011;12:407-30.
54. Desai AN, Jere A. Next-generation sequencing: ready for the clinics? Clin Genet. 2012;81(6): 503-10.

55. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu Y-M, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med. 2011;3(111):111-.
56. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11:49-79.
57. Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. Cancer Res. 1999;59(23):5975-9.
58. Auprich M, Bjartell A, Chun FKH, de la Taille A, Freedland SJ, Haese A, et al. Contemporary role of prostate cancer antigen 3 in the management of prostate cancer. Eur Urol. 2011;60(5):1045-54.
59. Sprenger CC, Plymate SR. The link between androgen receptor splice variants and castration-resistant prostate cancer. Hormones & cancer. 2014;5(4):207-17.
60. Antonarakis ES, Lu C, Wang H, Luber B, Nakazawa M, Roeser JC, et al. AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. The New England journal of medicine. 2014;371(11):1028-38.
61. Agarwal N, Di Lorenzo G, Sonpavde G, Bellmunt J. New agents for prostate cancer. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2014;25(9):1700-9.
62. Furgason JM, Bahassi el M. Targeting DNA repair mechanisms in cancer. Pharmacology & therapeutics. 2013;137(3):298-308.
63. Kaufman B, Shapira-Frommer R, Schmutzler RK, Audeh MW, Friedlander M, Balmana J, et al. Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(3):244-50.
64. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol. 2011;12(3):245-55.
65. Olmos D, Brewer D, Clark J, Danila DC, Parker C, Attard G, et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. Lancet Oncol. 2012;13(11):1114-24.
66. Ross RW, Galsky MD, Scher HI, Magidson J, Wassmann K, Lee G-SM, et al. A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study. Lancet Oncol. 2012;13(11):1105-13.
67. Casanova-Salas I, Rubio-Briones J, Fernandez-Serra A, Lopez-Guerrero JA. miRNAs as biomarkers in prostate cancer. Clin Transl Oncol. 2012;14(11):803-11.
68. Catto JWF, Alcaraz A, Bjartell AS, De Vere White R, Evans CP, Fussel S, et al. MicroRNA in prostate, bladder, and kidney cancer: a systematic review. Eur Urol. 2011;59(5):671-81.
69. Larne O, Martens-Uzunova E, Hagman Z, Edsjo A, Lippolis G, den Berg MSV-v, et al. miQ--a novel microRNA based diagnostic and prognostic tool for prostate cancer. International journal of cancer Journal international du cancer. 2013;132(12):2867-75.
70. Baca SC, Garraway LA. The genomic landscape of prostate cancer. Front Endocrinol (Lausanne). 2012;3:69-.

71. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J-P, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44(6):685-9.
72. Grasso CS, Wu Y-M, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012;487(7406):239-43.
73. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470(7333):214-20.
74. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell. 2010;18(1):11-22.
75. Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, et al. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med. 2010;16(7):793-8.
76. Goo YA, Goodlett DR. Advances in proteomic prostate cancer biomarker discovery. J Proteomics. 2010;73(10):1839-50.
77. Cocucci E, Racchetti G, Meldolesi J. Shedding microvesicles: artefacts no more. Trends Cell Biol. 2009;19(2):43-51.
78. Jansen FH, Krijgsveld J, van Rijswijk A, van den Bemd G-J, van den Berg MS, van Weerden WM, et al. Exosomal secretion of cytoplasmic prostate cancer xenograft-derived proteins. Mol Cell Proteomics. 2009;8(6):1192-205.
79. Duijvesz D, Luider T, Bangma CH, Jenster G. Exosomes as biomarker treasure chests for prostate cancer. Eur Urol. 2011;59(5):823-31.
80. Ronquist G. Prostasomes are mediators of intercellular communication: from basic research to clinical implications. J Intern Med. 2012;271(4):400-13.
81. Simpson RJ, Jensen SS, Lim JWE. Proteomic profiling of exosomes: current perspectives. Proteomics. 2008;8(19):4083-99.
82. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, Lotvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol. 2007;9(6):654-9.
83. Tavoosidana G, Ronquist G, Darmanis S, Yan J, Carlsson L, Wu D, et al. Multiple recognition assay reveals prostasomes as promising plasma biomarkers for prostate cancer. Proc Natl Acad Sci U S A. 2011;108(21):8809-14.
84. Duijvesz D, Versluis CYL, van der Fels CAM, Vredenbregt-van den Berg MS, Leivo J, Peltola MT, et al. Immuno-based detection of extracellular vesicles in urine as diagnostic marker for prostate cancer. International Journal of Cancer. 2015 in press.
85. Choudhury AD, Eeles R, Freedland SJ, Isaacs WB, Pomerantz MM, Schalken JA, et al. The role of genetic markers in the management of prostate cancer. Eur Urol. 2012;62(4):577-87.
86. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005;37(4):413-7.
87. Castro E, Eeles R. The role of BRCA1 and BRCA2 in prostate cancer. Asian J Androl. 2012;14(3):409-14.
88. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline mutations in HOXB13 and prostate-cancer risk. The New England journal of medicine. 2012;366(2):141-9.



Chapter 2

The *GPS2-MPP2* gene fusion promotes growth and decreases apoptosis in the LNCaP cell line

I. Teles Alves[1,2]
R. Böttcher[1]
M. van Royen[2]
N. Dits[1]
J. Trapman[2]
G. Jenster[1]

Departments of Urology[1] and Pathology[2],
Erasmus MC, Rotterdam, The Netherlands

Manuscript submitted

Abstract

Prostate cancer is a frequently diagnosed disease and the sixth leading cause of cancer-related death in men worldwide. It is a heterogeneous disease, both at the morphological and molecular level. Gene fusions are a frequent class of somatic alterations observed in these solid tumours with the *TMPRSS2-ERG* gene fusion being present in approximately 50% of all prostate cancer cases. In this study we have used SNP (single nucleotide polymorphism) and exon array data as well as RNA sequencing to identify novel gene fusions in prostate cancer. By combining these methods, the gene fusion *GPS2-MPP2* was detected in the LNCaP cell line. The *GPS2-MPP2* gene fusion causes the expression of a chimeric protein containing part of the coiled-coil domain of GPS2 encoded by the first exon of the gene, that is fused to the complete *MPP2* protein coding sequence except for the 1st L27 domain which is only partially encoded by the fusion gene (exon 2 and onwards). The specific knockdown of this chimeric protein resulted in a clear decrease of proliferation rate in LNCaP cells. Furthermore, knockdown also induced apoptosis with higher caspase 3/7 activation levels in LNCaP cells as compared to control cells. Live-cell imaging of LNCaP cells showed that the cellular localization of the GPS2-MPP2 fusion encompasses both the speckle-like and membrane-patch distribution of GPS2 and MPP2, respectively.

Introduction

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide (1). While many of these men will have an indolent form of the disease, others will develop an aggressive, life threatening condition (2). To date, PCa remains the sixth leading cause of cancer-related death among men. The spectrum of genetic alterations in PCa is wide and although a few genes are known to be frequently mutated in this disease, many have now been identified with low recurrence rate (3).

Gene fusions are a frequent class of somatic alterations in PCa (4). The overexpression of the ERG oncogene due to the TMPRSS2-ERG gene fusion is one of the earliest genetic alterations and occurs in approximately 50% of PCa cases (5). ERG is a member of the ETS family of transcription factors of which other members are also found to be fused in PCa, typically with prostate-specific and androgen-regulated fusion partners (6). Advances in high-throughput technologies have extended the number of gene fusions detected in prostate cancer (4, 7).

Novel rare gene fusions have been identified in PCa involving genes encoding members of the MAPK pathway KRAS, RAF1 and BRAF (8-10). In addition to being frequently overexpressed in PCa, cMYC is reported to be fused in frame with the gene partner C15orf21 (11). Also the tumour suppressor gene FH and members of the Rho guanine nucleotide exchange factor (GEF) family have been detected in gene fusions present in primary PCa as well as in lymph node metastases (11, 12).

The identification of novel gene fusions in PCa, generated limited insight into novel high frequency events, but it provided valuable information regarding individual tumour genomes (13). This is particularly important in the case of PCa due to its clinical and biological heterogeneity (4). In the end, these private gene fusions might aid in providing a distinct line of treatment for the patient such as the rearrangements in the RAF pathway or pinpoint to common molecular pathways to different rearrangements (14).

In this study we have used SNP and exon array data of a wide set of prostate cancer cell lines, xenografts and clinical PCa samples to detect novel gene fusions. We identified the in-frame gene fusion GPS2-MPP2 in the LNCaP cell line. The LNCaP PCa cell line is one of the most studied cell lines. It is derived from a biopsy of a supraclavicular lymph node of a patient with rapidly progressing PCa (15, 16). GPS2-MPP2 increases the proliferation capacity of LNCaP cells and protects against apoptosis. The cellular localization of GPS2-MPP2 mimicked both the cellular localization of GPS2 and MPP2 alone.

Results

Identification of novel gene fusions in prostate cancer

To identify novel DNA and RNA fusions in PCa, we used Illumina 1M SNP microarray data of 66 PCa samples and 131 normal control samples together with Affymetrix exon array data of 203 clinical PCa samples, 11 xenografts and 6 cell lines (see materials & methods). The SNP array data was processed with the Nexus software to list all the detectable DNA breaks. A total of 13,150 DNA breaks were identified in all the 66 PCa samples (Supplementary Table 1, see supplementary data. From all the breaks identified, 4589 overlapped with genes. Using the exon array data we were able to select which of the 4589 genes with breaks also showed abnormal expression.

To determine irregular expression patterns the transcript cluster probesets located on the left and right side of the DNA breakpoint were selected. For each sample, the FIRMA scores of these probesets were calculated and compared using Wilcoxon-testing to determine the significance of the differential expression of these genes. With the algorithm AltSplice Mapper we were able to visualize the probeset distribution of each abnormally expressed gene. For each sample, individual probesets were normalized to the expression of the whole transcript and to the corresponding probeset in all PCa samples.

A list with all the genes with DNA breaks and abnormal gene expression was generated (Supplementary Table 2, see supplementary data. We selected 3 candidates with the most prominent indication for abnormalities both at the DNA and RNA level. These 3 genes showed lower expression and a break in the beginning of the gene and were checked for possible gene partners by 5' gene RACE PCR (Rapid Amplification of cDNA Ends PCR) (Table 1).

Table 1: List of the top 3 candidate genes from the combined analysis of DNA and RNA data for detection of novel gene fusions. The event sample represents the sample in which the DNA break location corresponds to a shift in expression pattern. *MPP2* and *NRXN3* were successfully validated by 5´ race PCR. Chr, chromosome; ex, exon.

Gene	Chr	Break location (DNA)	Break location (RNA)	Event sample	Validated	Fusion partner	In frame
MPP2	17	39,334,282	ex 2-3	LNCaP	yes	*GPS2*	yes
NRXN3	14	78,246,877	ex 5-6	G295	yes	*TSHR*	no
YES1	18	779,852	ex 1-2	LAPC4	no	-	-

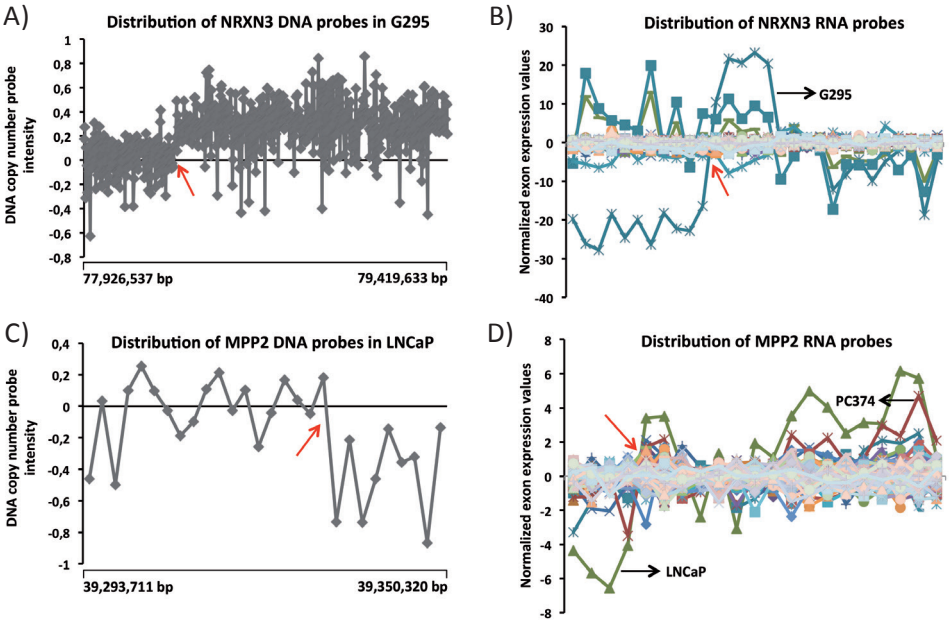


Figure 1: DNA and RNA breaks in the *NRXN3* and *MPP2* genes. The distribution of SNP probeset intensity values in the *NRXN3* (+ orientation) and *MPP2* (- orientation) gene are shown in (A) and (C), respectively. The shift between a positive and negative cluster of probeset values represents the breakpoint location at the genomic levels. The red arrow depicts the location of the break. The genomic region shown, contains all the probesets within the gene together with five extra probe locations upstream and downstream of the gene. The distribution of exon expression values for the *NRXN3* and *MPP2* transcripts (left is 5'; right is 3') are shown in (B) and (D), respectively. Only core exon probeset values are displayed. The red arrow shows where the DNA break is in the exon array plots. In (B), the G295 sample shows a relative lower expression of the first as compared to the last exon probesets. The expression of the first exon probesets is also significantly lower compared to the average expression of the same probesets in the other PCa samples. In (D), the distribution of the *MPP2* exon probeset values is displayed. The expression values of the first exon probesets are lower compared to the expression values of the last exon probesets.

Validation of novel gene fusions in prostate cancer

We were able to identify and validate a fusion partner for 2 out of the 3 genes. First, we identified *TSHR* (thyroid stimulating hormone receptor) by 5' RACE as a fusion partner of *NRXN3* (neurexin 3) in the G295 patient sample (Figure 1A and 1B). We validated the *TSHR-NRXN3* gene fusion by cDNA Sanger sequencing of RNA extracted from the G295 patient sample. This gene fusion was not in-frame and since it would not result in a chimeric protein, it was excluded from further analysis. We identified GPS2 (G protein pathway suppressor 2) as a fusion partner of MPP2 (membrane protein, palmitoylated 2 (MAGUK p55 subfamily member 2) in the LNCaP cell line (Figure 1C and 1D). By Sanger sequencing we validated the GPS2-MPP2 fusion in LNCaP. This fusion was predicted to

be in-frame resulting in the expression a chimeric protein. Both *GPS2* and *MPP2* are located in chromosome 17. *GPS2* is located in 17p whereas *MPP2* is located in 17q. *GPS2* encodes a protein involved in MAPK signalling cascades and *MPP2* is a member of the membrane palmitoylated subfamily of proteins with important roles in cellular signal transduction. The *GPS2-MPP2* fusion gene was selected to further investigate the potential consequences of the gene fusion.

GPS2-MPP2 gene fusion in the LNCaP cell line

The *GPS2-MPP2* fusion encompasses, at the transcript level, exons 1 and 2 of *GPS2* spliced to exon 2 of *MPP2* (canonical isoform) (Figure 2A). This resulted in the expression of a chimeric protein with part of the coil-coiled domain of *GPS2* and a large part of *MPP2*, with exception of the 1st L27 domain, which is only partially present (Figure 2B). The fusion was also observed in RNAseq data of the LNCaP cell line (data not shown) (17). We performed Fluorescence *in situ* hybridization (FISH) assays of *GPS2* and *MPP2* in LNCaP, PNT2C2 and normal control metaphase spreads and confirmed the colocalization of *GPS2* and *MPP2* in LNCaP nuclei. On average, there were 4 copies of both *GPS2* and chromosome 17 and between 5 and 10 copies of *MPP2* present in LNCaP cells (Figure 2C). Regarding *GPS2* and chromosome 17 this was expected given the ~4N ploidy of this cell line determined previously by spectral karyotyping (18). Typically, only one of the 4 copies co-localized, which was observed in 85% of LNCaP cells counted. In the immortalized normal prostate epithelial PNT2C2 cells there were between 2 and 5 chromosome 17 centromere probe signals, 3 and 6 *GPS2* probe signals and 4 and 10 *MPP2* probe signals. Approximately 10% of the nuclei showed colocalization of *GPS2* and *MPP2* signals. In the control sample (normal blood cell nuclei) the majority of nuclei displayed only two FISH signals for *GPS2, MPP2* and the centromeric probe 17. The percentage of co-localization of *GPS2* and *MPP2* signals was 2%.

RT-PCR of RNA from 201 PCa patient samples did not reveal the presence of the fusion transcript in any sample. Although the expression pattern of *MPP2* in the PC133 and PC374 PCa xenografts followed the same trend as that observed in the LNCaP sample (Figure 1D), the presence of the *GPS2-MPP2* fusion could not be confirmed at the RNA level. Overall expression levels of *MPP2* did not show an outlier expression for LNCaP harbouring the *GPS2-MPP2* gene fusion (Supplementary Figure 2). Therefore, upregulation of *MPP2* expression levels does not explain the positive selection for this fusion in LNCaP cells as the explanation is in case of gene fusions involving ETS genes.

Additionally, average *MPP2* and *GPS2* levels were significantly higher in normal samples as compared to PCa samples (Supplementary Figure 1 and 2). Also *MPP2* levels were significantly lower in lymph node metastases and transurethral resections of the prostate (TURP) as compared to the primary tumor. So far, there is limited knowledge regarding the regulation of *GPS2* and *MPP2* expression and no reliable antibodies are available to study the *GPS2* and *MPP2* proteins.

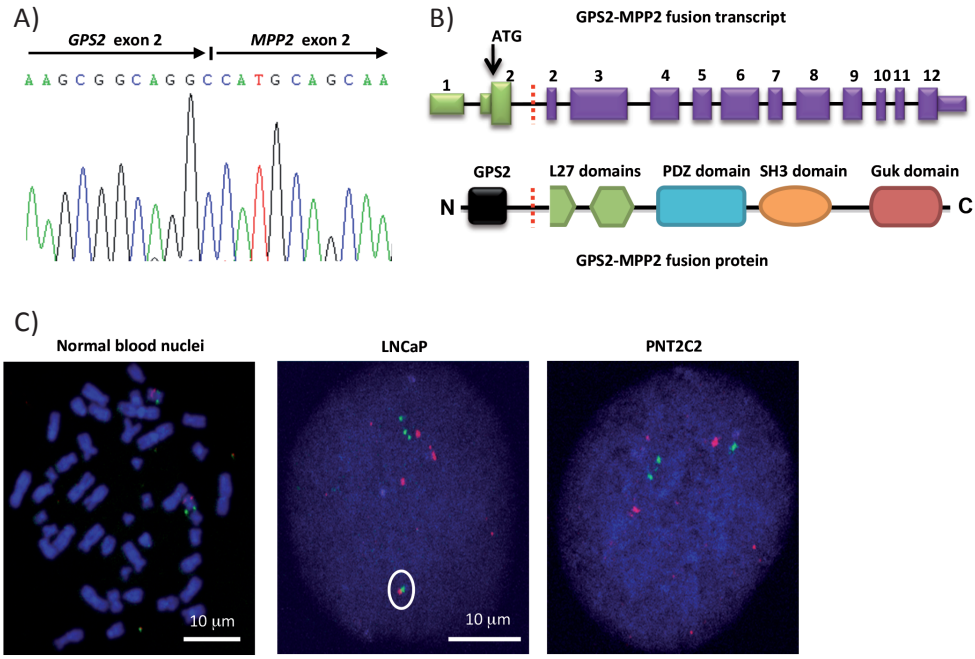


Figure 2: Validation of the *GPS2-MPP2* gene fusion. The *GPS2-MPP2* gene fusion was sequenced in the LNCaP cell line (A). We confirmed the fusion of exon 2 of *GPS2* to exon 2 of *MPP2*. The scheme of the *GPS2-MPP2* fusion transcript and chimeric protein is depicted in (B). In (C), the distribution of FISH probe signals is displayed, with *MPP2* and *GPS2* corresponding to the red and green signal, respectively. The number of signals and signal colocalization was scored per nuclei. For each sample a total of 45 nuclei were assessed. 85% of the LNCaP nuclei showed colocalization of *GPS2* and *MPP2* probe signals as opposed to 10% and 2% in PNT2C2 nuclei and normal control metaphase spreads, respectively.

Cellular localization of the GPS2-MPP2 protein differs from GPS2 and MPP2

We investigated the cellular localization of *GPS2, MPP2* and the *GPS2-MPP2* chimeric protein, LNCaP cells were transiently transfected with a vector expressing *GPS2, MPP2* or *GPS2-MPP2* fused in-frame to *TurboGFP*. Live cell imaging was performed for all the different tagged proteins (Supplementary Movie 1-4, see supplementary data . GPS2 showed a speckle-like distribution mainly in the cytoplasm and to a lesser extend in the nucleus (Figure 3A). These speckles overlapped with PKH26 red labelled vesicles although colocalization could not be observed for all speckles in a given cell (Figure 3B). *MPP2* displayed a faint overall cytoplasmic distribution with brighter green membrane patches (Figure 3C). These patches were not present in all cells. Regarding the *GPS2-MPP2* chimeric protein, we often observed a mixed distribution with cells showing a more speckle-like pattern or cells with membrane patches and with cells with both speckles and membrane patches (Figure 3D).

The GPS2-MPP2 chimeric protein promotes proliferation and protects against apoptosis

To determine the functional consequences of the *GPS2-MPP2* gene fusion, we designed siRNA oligonucleotides specifically targeting the *MPP2* transcript and the *GPS2-MPP2* transcript. The knockdown using siRNA against *MPP2* strongly affected *MPP2* expression in all cell lines tested (Supplementary Figure 3). The siRNA directed against the *GPS2-MPP2* fusion transcript was only effective in downregulating *GPS2-MPP2* expression without affecting the *MPP2* transcript levels (Supplementary Figure 3).

We performed a proliferation assay of LNCaP cells transfected with the siRNA against MPP2 alone and a scrambled control siRNA. A significant decrease in proliferation capacity in *MPP2* knockdown cells as compared to control cells was observed (Figure 4A). To investigate whether this phenotype was a result of the downregulation of the *GPS2-MPP2* gene fusion, we tested a siRNA solely targeting the *GPS2-MPP2* fusion transcript. Again, we observed a significant decrease in proliferation rate of LNCaP cells, in this case due to *GPS2-MPP2* knockdown, as compared to control cells (Figure 4B).

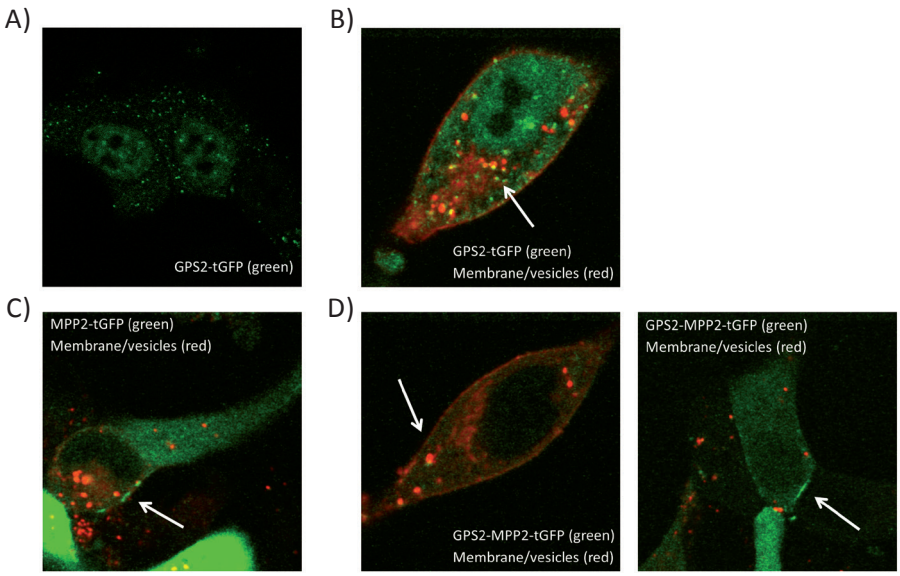


Figure 3: Cellular localization of *GPS2, MPP2* and the *GPS2-MPP2* proteins. In (A), GPS2-tGFP shows a speckle-like distribution across the cytoplasm with some nuclear staining. In (B), live cell imaging of GPS2-tGFP is combined with the red cellular membrane staining PKH26 to examine colocalization of the GPS2-tGFP speckles with membrane-derived vesicles. The white arrows point to colocalization. In (C), MPP2-tGFP shows overall cytoplasmic staining with specific locations where membrane patches are strongly positive. In (D), the GPS2-MPP2-tGFP protein displays a mixed cellular localization pattern with both speckles that coincide with membrane-derived vesicles and membrane patches (indicated by white arrows).

The same knockdown experiment was performed using the RWPE-1 immortalized normal prostate epithelial cell line with no effect on the proliferation capacity of these cells (Figure 4C).

In addition to the effect of *GPS2-MPP2* on cellular proliferation, we addressed the influence of this fusion on cellular apoptosis. The PNT2C2 normal prostate epithelial cell line was used as a control cell line instead of RWPE-1 due to the poor performance of this cell line in the measurement of Caspase 3/7 levels. Both PNT2C2 and LNCaP cells were transfected with control and *GPS2-MPP2* targeting siRNAs. Caspase 3/7 activity levels were elevated in LNCaP cells with *GPS2-MPP2* knockdown as compared to the control PNT2C2 cells (Figure 4D). We concluded that the presence of the *GPS2-MPP2* gene fusion induces cellular proliferation as well as a protective effect against apoptosis.

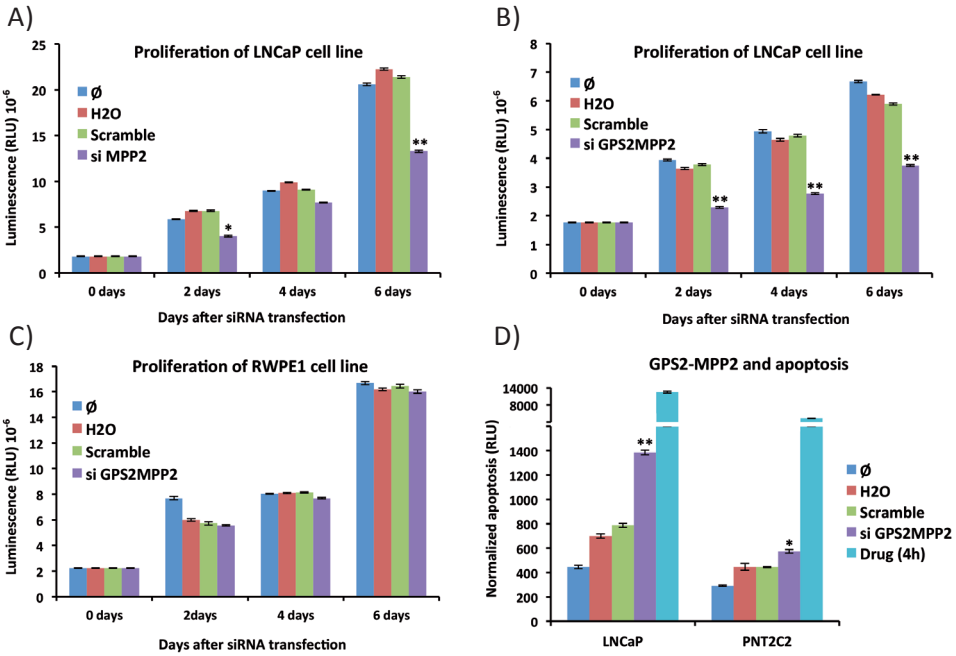


Figure 4: *GPS2-MPP2* promotes cell proliferation and protects against apoptosis in LNCaP cells. In (A), the proliferation of LNCaP cells was measured after *MPP2* knockdown. The ø represents LNCaP cells growing under normal conditions, H2O the addition of transfection reagent, scramble the addition of a non-targeting small interfering RNA and siMPP2 the knockdown of the *MPP2* transcript. An asterisk * P ≤0.02 and **P ≤ 0.002 by a two-tailed Student's t-test. In (B), the cell proliferation of LNCaP cells was measured after *GPS2-MPP2* knockdown. In (C), the cell proliferation of RWPE1 cells was measured after *MPP2* knockdown. In (D), apoptosis was measured through the detection of Caspase 3/7 activity corrected for cell viability. This assay was performed for both LNCaP and PNT2C2 cell lines. The same controls of the proliferation assays were used along with an apoptosis control (drug-treated cells for 4 h).

Discussion

Gene fusions are common genetic abnormalities found in PCa (19). So far, several gene fusions involving members of the ETS family of transcription factors have been identified (20). *TMPRSS2-ERG* is present in approximately 50% of all prostate cancers whereas fusions with other members of the ETS family are less frequent (14). In our study we used SNP and exon array data to identify novel gene fusions in PCa. The combination of DNA and RNA information allows the exclusion of DNA abnormalities that do not alter expression patterns. In this way, we expect to find abnormalities that will most likely have functional implications to the cancer cells.

One of the validated gene fusions involved the genes *NRXN3* and *TSHR. NRXN3* is one of the three neurexin (NRXN) genes known in humans (21). Despite the clear specificity of this gene to neuronal-related functions, other *NRXN3* gene isoforms have been shown to play a role in cardiac intercellular connections and in pancreatic beta cells (22, 23). The fusion of *NRXN3* to *TSHR* does not result in an in-frame chimearic protein but could produce a shorter *NRXN3* protein due to the potential use of an alternative start codon in the *NRXN3* transcript. *TSHR* encodes the thyroid stimulating hormone (TSH) receptor and mutations in this gene have been identified in tumours of the thyroid gland. Although the relevance of this *NRXN3* isoform in PCa cannot be discarded, the low expression levels of *TSHR* in G295 will result in low abundance of this transcript.

LNCaP is a cell line derived from a lymph node metastasis of a PCa patient (24). With the SNP array data we were able to detect a break at the DNA level in the gene *MPP2*. The expression pattern of *MPP2* showed lower expression of the first exon of *MPP2* and a relative overexpression of the remaining exons consistent with the break location at the genomic level. Using 5' RACE PCR we determined *MPP2* to be fused to *GPS2*. The *GPS2-MPP2* gene fusion results in a chimeric protein containing part of the coiled-coil domain of *GPS2* fused to part of the 1st L27 domain and remaining domains of *MPP2*.

GPS2 is a suppressor of G protein-activated MAPK signalling (25). This 37 kDa protein possesses an N-terminal coiled-coil domain and an α-helix motif in the N-terminus which is responsible for the interaction with Transducin (Beta)-Like 1 (TBL1), a protein involved in transcription regulation (26). This sequence motif is conserved in both SMRT and its homolog N-COR and in all orthologues of SMRT, N-COR and *GPS2* (26). *GPS2* has been indicated by several studies to play a role in transcriptional regulation (27). In addition, *GPS2* is also involved in DNA repair (28), metabolism (29), cytoskeleton architecture (30), inflammation (27) and cell cycle regulation (31). Although many of the functions known to *GPS2* focus on its transcriptional repressor capacity, this protein has also been found to enhance transactivation of several transcription factor such as p53 (32, 33). In addition, *GPS2* is also found to interact with TNFR1, revealing cytoplasmic functions of *GPS2* in the regulation of both the JNK and the TNF-α proinflammatory pathways (27).

MPP2 is a member of the MPP subfamily of the Membrane-Associated Guanylate Kinase (MAGUK) family of proteins (34). This family consists of ten subfamilies and a total of 22 members with variations in both size and domain organization (35). The conserved core structure of this family consists of one to multiple PDZ domains, a Src homology 3 (SH3) and a catalytically inactive guanylate kinase (GK) domain (36) (Supplementary figure 4). The GK domain, although catalytically inactive, still functions as a specific phosphopeptide binding module (37). Members of the *ZO 2-3* subfamily have been shown to transiently locate to the nucleus and interact with a number of nuclear proteins (38). Members of the MAGUK family of proteins, involved in synaptic regulation, are already considered for pharmaceutical intervention in the case of neurological diseases (39). Although some MPP members have already been linked to cancer, namely *MPP3* in hepatocellular carcinoma (40), a link between *MPP2* and cancer was not yet established. MPP2 is the human homolog of the tumour suppressor protein DLG2 in Drosophila. Despite being a tumour suppressor this protein has also shown oncogenic functions dependent of the precise cellular context (41).

The identification of a novel gene fusion between *GPS2* and *MPP2* expands the currently knowledge regarding these two proteins. *MPP2* is the second MPP family member to be rearranged in PCa. Previously, we independently identified the functional and in-frame MPP5-FAM71D fusion in PC346C cells (12). Regarding *GPS2*, only one report so far has described mutations in this gene in medulloblastoma (42). In the *GPS2-MPP2* gene fusion we have a chimeric protein with the first 31 aa of *GPS2* fused to aa 12 to 552 of MPP2. The first 31 aa of GPS2 contain the N-terminal TBL1-binding α-helix motif and aa 12 to 552 of *MPP2* lack 4 aa from the first L27 domain of *MPP2* (L27N). L27 domains are common to many scaffold proteins and are involved in the assembly of supramolecular complexes that participate in cell polarity (43). Despite the membrane localization, there is no signal peptide in the N-terminal region of *MPP2*. So far, studies suggest that MAGUK proteins containing tandem L27 domains such as the *MPP2* protein require both domains for the formation of ternary complexes (44). Nevertheless, the 4 aa residues missing from the L27N domain do not seem to be essential for the molecular organization of the L27N domain itself (45).

To address the properties of this novel chimeric protein, we determined the cellular localization of *GPS2, MPP2* and the *GPS2-MPP2* fusion. The latter showed a combined phenotypic distribution of both *GPS2* and *MPP2* alone. Although it is difficult to predict what the consequence of this distribution pattern might be on protein function, we can assume that it will likely have an effect on cellular pathways regulated by the fusion product. Due to the fusion, *MPP2* is found in vesicle-like structures as well as in the membrane and part of *GPS2* will consequently also localize now to the membrane.

The knockdown of the *GPS2-MPP2* fusion in LNCaP cells affected both proliferation and apoptosis. LNCaP cells showed a decreased proliferation after the knockdown of the fusion as compared to control cells. The knockdown of *MPP2* alone in RWPE1

normal epithelial prostate cells did not cause any detectable alteration in proliferation capacity. Caspase 3/7 activation was greatly increased in LNCaP cells following *GPS2-MPP2* knockdown. This effect was not observed in PNT2C2 normal epithelial prostate cells, suggesting a role of the *GPS2-MPP2* gene fusion in protecting against apoptosis. Although members of the MAGUK family of proteins have already been implicated in the regulation of cell proliferation and apoptosis (46-48), so far no report has shown this for the MPP subfamily.

Currently, most studies use next generation sequencing (NGS) technologies to identify novel gene fusions. The major advantage of these technologies is the unbiased approach, high resolution and ability to detect more than the known spectrum of genes. Nevertheless, microarray technology has been widely used for the past 20 years with the benefits that sample preparation is easier and data analysis tools are well established. Also, the limitations and pitfalls of microarray technologies are quite well known, which is not yet fully the case for NGS. Whether microarrays or NGS technologies are the suitable option will depend on the aims of the study. The detection of novel gene fusions is better facilitated by the use of NGS but microarray technologies remain a feasible option for validation of known fusions, especially when there is a considerable number of samples to be analyzed.

In conclusion, we show in this study the presence of a novel gene fusion *GPS2-MPP2* in PCa. Although this fusion seems to be specific for LNCaP, also *MPP5* was found to be rearranged in PCa (12). This fact suggests that molecular pathways linked to both these genes most likely play a role in cancer. Whether *MPP2* itself also displays this tumour suppressor/oncogene duality remains to be proven. Regarding *GPS2*, the multiplicity of pathways and functions targeted by this protein make it a desirable target for mutation in cancer. Nevertheless, only in melanoma this link has been shown. The *GPS2-MPP2* gene fusion here identified showed oncogenic potential, by promoting growth and suppressing apoptosis at the same time. Additional studies will help determine the exact pathway and mechanism, which links these genes with cancer.

Materials and Methods

SNP array data

Fresh-frozen clinical samples were obtained from the tissue bank of the Erasmus University Medical Center. Patient sample collection was performed according to national legislation concerning ethical requirements. Erasmus MC Medical Ethics Committee according to the Medical Research Involving Human Subjects Act (MEC-2004-261) has approved the use of these samples.
Illumina 1M SNP Array analysis from 47 PCa patients, 11 PCa xenografts, 8 PCa cell lines and 6 non-cancer patients were collected. Genotyping was performed using the Infinium Illumina Human 1M probe BeadChip containing 1,072,820 markers of which

206,665 are in CNV regions. The experiments were performed by ServiceXS (Leiden, The Netherlands) and within the group of André Uitterlinden (Erasmus MC) according to the manufacturer's protocol (Illumina) (49).
Nexus Copy Number 5.0 (BioDiscovery) was used for data analysis. The Human 1M CV HapMap control set provided by the manufacturer was used as a control. SNP-FASST Rank segmentation was the algorithm used for sample analysis. Standard settings were adjusted to: significant threshold - 10^{-7}, maximum contagious probe spacing - 1000 kb, minimal number of probes per segment - 15, high gain - 0.6, gain -0.2, loss - -0.2, big loss - -1, homozygous frequency threshold - 0.9, homozygous threshold - 0.85, heterozygosity imbalance threshold - 0.35 and minimum loss of heterozygosity (region with loss of heterozygosity) length - 5000 (kb).

Affymetrix exon arrays

Exon-array data from 89 PCa patient samples, 11 PCa xenografts and 6 cell PCa cell lines were collected using the GeneChip Human Exon 1.0 ST Array (Affymetrix). Staining, washing and scanning procedures were performed according to the manufacturer's protocol (Affymetrix). Processing and RMA quantile normalization of the data were performed using the R-package affy (http://www.bioconductor.org/packages/release/bioc/manuals/affy/man/affy.pdf).
The list of genes with DNA breakpoint originated from the SNP microarray data was analyzed to determine differential expression of exons in these genes (GSE41410). On the GeneChip Human Exon 1.0 ST Array, 5,362,207 probes (on average 40 probesets per gene) are used to analyze one million exon clusters (collections of overlapping exons). The file containing the raw intensity values for the core probes (derived from RefSeq transcripts or full-length mRNAs) was used to assess exon expression profile.

Integration of Illumina SNP arrays and Affymetrix Human Exon Arrays to identify potentially transcribed gene fusions

For 54 SNP arrays used in this study, matching expression data from Affymetrix Human Exon Arrays were available (17, 49), which were normalized using the RMA implementation of the aroma.affymetrix R-package and further processed using the FIRMA method (50, CDF used: HuEx-1_0-st-v2,fullR3,A20071112,EP.CDF, see http://www.aroma-project.org/). In short, FIRMA was designed to identify alternative splicing events from Human Exon Arrays by comparing the expected gene expression profile to the profiles of each of its exons and scoring whether systematic deviations can be found. Since gene fusions may cause parts of a gene to be expressed, the observed change in expression would mimic alternative splicing events measured by the array and therefore can be identified by using the same methodology.
To find breakpoints located within genes, we defined the start and end positions of copy number segments called by Nexus as genomic breaks (file 'cncalldetails.txt) and overlapped them with the RefSeq gene annotation (build hg18 available at http://genome.ucsc.edu/). After translating the RefSeq gene identifiers to gene symbols, Affymetrix transcript clusters that represented the candidate genes were defined. Next,

all corresponding 'core' and 'extended' probesets that were located on the left and right side of the DNA breakpoint were selected, respectively. For each sample, we then tested whether the FIRMA scores of probesets located before and after the breakpoint differed using a Wilcoxon-test.

Public array datasets

The prostate cancer dataset mentioned contains 48 previously published prostate cancer samples (GSE41408, (49)) as well as additional cancerous and control samples, accessible via GEO accession number GSE59745 (17). These datasets comprise samples from normal adjacent prostate (NAP), localized prostate cancer obtained via radical prostatectomy (PCa) and transurethral resection of the prostate (TURP), as well as metastasis in lymph node (LN PCa).

RNA and DNA isolation

Total RNA was isolated using the RNeasy kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol. Concentration and purity of RNA were assessed using the NanoDrop ND1000 spectrophotometer (Nanodrop products, Wilmington, DE, USA). RNA was stored at -80 degrees. DNA was isolated using the QIAamp DNA Blood Midi Kit (Qiagen) according to the manufacturers' instructions. Concentration and purity of the DNA were assessed using the NanoDrop ND 1000 spectrophotometer (Nanodrop products) and stored at -20 degrees.

Q-PCR

mRNA expression was analysed by quantitative PCR (qPCR). Complementary DNA was prepared with MMLV-RT (Invitrogen) and oligo(dT)12 primer according to manufacturer's protocol. Taqman probe Hs01115789_m1 was used to detect MPP2 expression levels. The detection of GPS2-MPP2 was done using a costum Taqman gene expression assay, with the 5' GCA AGC GGC AGG CCA TGC AGC A 3' oligonucleotide probe and primers 5' GCA CAT TAT GAT GGA GCG GGA 3' (Fw) and 5' GAG GCA TTA TGG AAA GTC CCA 3' (Rv). For detection of GAPDH the commercial taqman probe Hs02758331_g1 was used.

Cell culture

RWPE-1 and PNT2C2 (kindly provided by Prof. Norman Maitland, York, U.K.) are normal prostate epithelium cell lines. RWPE-1 cells were cultured in keratinocyte medium (GIBCO), supplemented with 5 ng/ml epidermal growth factor, 1% penicillin–streptomycin (BioWhittaker) and 50 mg/l bovine pituitary extract. LNCaP and PNT2C2 cells were cultured in RPMI-1640 Medium (GIBCO) supplemented with 10% fetal bovine serum and 1% penicillin–streptomycin (BioWhittaker). All cell lines were cultured at 37 degrees with 5% carbon dioxide.

RNA ligase-mediated rapid amplification of cDNA ends

5’ RACE was performed using the GeneRacer kit (Invitrogen) according to the manufacturer’s protocol. Briefly, cDNA was amplified using the GeneRacer 5’ primer and a reverse gene-specific primer (*MPP2* – exon 4; *NRXN3* – exon 6; *YES1* – exon 4). The PCR products were analysed on a 1% agarose gel and bands were excised, purified using a gel extraction kit (Qiagen) and sequenced.

Sequencing

Purified PCR products have been sequenced bidirectionally using standard Sanger sequencing. Sequencing PCR was carried out using the same reverse primer as the primer used in the 5’ RACE PCR. The forward primer was later designed, according to the gene partner identified. Primer concentrations were 3 ng of primer per reaction in a 20 µl reaction volume. The PCR product was sequenced on an ABI Model 3730 automated sequencer and analyzed using DNAMAN (Lynnon Corporation, Pointe-Claire, QC, Canada).

FISH

FISH was performed on nuclei spreads from LNCaP, PNT2C2 and normal blood cells. Cells were collected and incubated at 37 degrees Celsius for 18 minutes in a 0.075M KCL solution. Fixation was then performed using a freshly made 3:1 Methanol/ice-acetic acid solution (fixative). Cells were centrifuged for 5 min at 200 g. Cells were resuspended in 5 ml of fixative and centrifuged for 5 min at 200 g. This step was performed 3 times. Cells were dropped onto the slide and allowed to dry for 20 min. BAC clones RP11-1099M24 (*GPS2*, before the break) and RP11-290H9 (MPP2, after the break) were purchased from BacPac Resources (bacpac.chori.org) and SE 17 (D17Z1) centromere probe 17 was purchased from Kreatech Diagnostics (Netherlands). BACs were labeled with digoxigenin-11-dUTP or biotin-16-dUTP (Roche) and visualized with anti-digoxigenin FITC (Roche) or streptavidin-Alexa 594 (Invitrogen). Labelling was performed according to the manufacturer’s protocol. FISH was performed according to standard protocols with minor modifications (51).

GPS2, MPP2 and *GPS2-MPP2 TurboGFP* constructs

The sequence validated I.M.A.G.E clones IRATp970F0313D (*GPS2*) and IRATp970C0650D (*MPP2*) corresponding to the BC013652 and BC030287 clone accession numbers were purchased from Source Bioscience, Nottingham, UK. Primers 5’ ATG CAG CCA GCA GGA AAG TC 3’ (Fw) and 5’ CTC CCG CTC CAT CAT AAT GT 3’ (Rv) and 5’ ATG CCT CGA AGG AAG ATC AG 3’ (Fw) and 5’ CTG CCT GGT CAA TAG CAA CCT 3’ (Rv) were used to amplify the *GPS2* and *MPP2* cDNA respectively and subsequently clone the fragment in frame into the Mammalian expression vector pTurboGFP-C (Evrogen). *GPS2-MPP2* cDNA was synthesized by GENECUST EUROPE (Luxembourg) and cloned into the pUC57 plasmid. The *GPS2-MPP2* cDNA was cut using XbaI and BamHI and subsequently ligated

to pTurboGFP-C. The amplified fragments were ligated to the vector using the Rapid DNA Ligation Kit according to the manufacturer's protocol (Roche). JM109 competent bacteria (Promega) were transformed with the ligation products using the heat shock method and plated in Kanamycin LB agar plates. DNA isolation was performed on selected colonies (Qiagen) and further sequence to confirm the cloning products.

Live-cell imaging

LNCaP and RWPE1 cells were cultured in coverslips at day 0 and transiently transfected at day 1 with *TurboGFP-GPS2, TurboGFP-MPP2* and *TurboGFP-GPS2-MPP2* using Fugene 6 (Promega) according to the manufacturer's protocol. After 48 h cells were imaged using the LSM 510 NLO confocor II (Zeiss). In addition, cells transfected with *TurboGFP-GPS2* and *TurboGFP-GPS2-MPP2* were also stained with the PKH26 Red Fluorescent Cell Linker for General Cell Membrane Labeling (Sigma) according to standard protocol. A time-lapse was performed for *TurboGFP-GPS2* and *TurboGFP-GPS2-MPP2*.

Functional assays

The functional role of *GPS2-MPP2* in the LNCaP cell line was assessed with the knockdown of both *MPP2* and *GPS2-MPP2* fusion gene. Proliferation assays were performed at different time points. The small interfering RNA reagent siGENOME SMARTpool, Human *MPP2* (ThermoFisher Scientific) interacts with a part of the *MPP2* gene, which is also present in the fusion. Briefly, LNCaP and RWPE-1 (control) cells were plated in 96-well culture plates at a density of 2000 cells per well (100 µl) and plated in T25 tissue flasks. Outer wells of the culture plates were filled with phosphate-buffered saline in order to avoid culturing artifacts in the proliferation assays. After 2 days, cells were washed using phosphate-buffered saline and transfected with Human *MPP2* (siGENOME SMARTpool) or the custom *GPS2-MPP2* siRNA oligos (sense GGC AGG CCA UGC AGC AAG UdT dT and anti-sense ACU UGC UGC AUG GCC UGC CdT dT) (Sigma) (final concentration of 100 nM) and Dharmafect3 reagent according to the manufacturer's protocol (Thermo Scientific) in medium without P/S. As negative controls, medium only, Dharmafect3 with water and a transfection with non-targeting small interfering RNA were included. Twenty-four hours after transfection the medium was replaced with normal medium, without P/S and cells were allowed to grow for a maximum of 6 days. Proliferation was assessed using the Cell Titer-Glo Luminescent Cell Viability Assay (Promega) according to the manufacturer's protocol. Apoptosis was assessed using the ApoLive-Glo Multiplex Assay (Promega) that measures both the number of viable cells as a marker of cytotoxicity and caspase activation as a marker of apoptosis within a single assay well. The assay was performed according to the manufacturer's protocol. As positive control of apoptosis an apoptosis inducer set (Millipore) was used. A mix containing 700x dilutions of Actinomycin D (10 mM), Camptothecin (2 mM), Cycloheximide (100 mM), Dexamethasone (10 mM) and Etoposide (10 mM) was used for 4 hours to induce apoptosis in both LNCaP and PNT2C2 cell lines.

Acknowledgements

We would like to thank André Uitterlinden from the Department of Internal Medicine and the Center for Biomics, Erasmus MC for microarray assistance, Arno van Leenders from the Department of Pathology, Erasmus MC for patient sample selection, Bert Eussen and Annelies de Klein from the Department of Clinical Genetics, Erasmus MC for the assistance with SNP microarray analysis and Service XS, Leiden and Barry van der Mast for microarray and software assistance. This research was made possible by financial contributions from the FP7 Marie Curie Initial Training Network PRO-NEST (grant number 238278) and the Foundation for Scientific Urological Research (SUWO).

References

1. Jemal A, Center MM, DeSantis C, Ward EM. Global Patterns of Cancer Incidence and Mortality Rates and Trends. Cancer Epidemiology Biomarkers & Prevention. 2010;19(8):1893-907.
2. Barbieri CE, Bangma CH, Bjartell A, Catto JW, Culig Z, Gronberg H, et al. The mutational landscape of prostate cancer. European urology. 2013;64(4):567-76.
3. Hessels D, Schalken JA. Recurrent gene fusions in prostate cancer: their clinical implications and uses. Curr Urol Rep. 2013;14(3):214-22.
4. Wyatt AW, Mo F, Wang Y, Collins CC. The diverse heterogeneity of molecular alterations in prostate cancer identified through next-generation sequencing. Asian journal of andrology. 2013;15(3):301-8.
5. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusionof TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005;310(5748):644-8.
6. Gasi Tandefelt D, Boormans J, Hermans K, Trapman J. ETS fusion genes in prostate cancer. Endocrine-related cancer. 2014.
7. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome research. 2011;21(1):56-67.
8. Beltran H, Yelensky R, Frampton GM, Park K, Downing SR, MacDonald TY, et al. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. European urology. 2013;63(5):920-6.
9. Bakin RE, Gioeli D, Sikes RA, Bissonette EA, Weber MJ. Constitutive activation of the Ras/mitogen-activated protein kinase signaling pathway promotes androgen hypersensitivity in LNCaP prostate cancer cells. Cancer research. 2003;63(8):1981-9.
10. Palanisamy N, Ateeq B, Kalyana-Sundaram S, Pflueger D, Ramnarayanan K, Shankar S, et al. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nature medicine. 2010;16(7):793-8.
11. Wu C, Wyatt AW, Lapuk AV, McPherson A, McConeghy BJ, Bell RH, et al. Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. The Journal of pathology. 2012;227(1):53-61.
12. Teles Alves I, Hartjes T, McClellan E, Hiltemann S, Bottcher R, Dits N, et al. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. Oncogene. 2014.
13. Nakagawa H. Prostate cancer genomics by high-throughput technologies: genome-wide association study and sequencing analysis. Endocrine-related cancer. 2013;20(4):R171-81.
14. White NM, Feng FY, Maher CA. Recurrent rearrangements in prostate cancer: causes and therapeutic potential. Current drug targets. 2013;14(4):450-9.
15. Spans L, Helsen C, Clinckemalie L, Van den Broeck T, Prekovic S, Joniau S, et al. Comparative genomic and transcriptomic analyses of LNCaP and C4-2B prostate cancer cell lines. PloS one. 2014;9(2):e90002.
16. Spans L, Atak ZK, Van Nieuwerburgh F, Deforce D, Lerut E, Aerts S, et al. Variations in the exome of the LNCaP prostate cancer cell line. The Prostate. 2012;72(12):1317-27.

17. Bottcher R, Hoogland AM, Dits N, Verhoef EI, Kweldam C, Waranecki P, et al. Novel long non-coding RNAs are specific diagnostic and prognostic markers for prostate cancer. Oncotarget. 2015;6(6):4036-50.
18. Beheshti B, Karaskova J, Park PC, Squire JA, Beatty BG. Identification of a high frequency of chromosomal rearrangements in the centromeric regions of prostate cancer cell lines by sequential giemsa banding and spectral karyotyping. Mol Diagn. 2000;5(1):23-32.
19. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nature reviews Cancer. 2008;8(7):497-511.
20. Dean M, Lou H. Genetics and genomics of prostate cancer. Asian journal of andrology. 2013;15(3):309-13.
21. Ichtchenko K, Nguyen T, Sudhof TC. Structures, alternative splicing, and neurexin binding of multiple neuroligins. J Biol Chem. 1996;271(5):2676-82.
22. Occhi G, Rampazzo A, Beffagna G, Antonio Danieli G. Identification and characterization of heart-specific splicing of human neurexin 3 mRNA (NRXN3). Biochem Biophys Res Commun. 2002;298(1):151-5.
23. Suckow AT, Comoletti D, Waldrop MA, Mosedale M, Egodage S, Taylor P, et al. Expression of Neurexin, Neuroligin, and Their Cytoplasmic Binding Partners in the Pancreatic β-Cells and the Involvement of Neuroligin in Insulin Secretion. Endocrinology. 2008;149(12):6006-17.
24. Horoszewicz JS, Leong SS, Chu TM, Wajsman ZL, Friedman M, Papsidero L, et al. The LNCaP cell line--a new model for studies on human prostatic carcinoma. Progress in clinical and biological research. 1980;37:115-32.
25. Jin D-Y, Teramoto H, Giam C-Z, Chun RF, Gutkind JS, Jeang K-T. A Human Suppressor of c-Jun N-terminal Kinase 1 Activation by Tumor Necrosis Factor α. Journal of Biological Chemistry. 1997;272(41):25816-23.
26. Oberoi J, Fairall L, Watson PJ, Yang JC, Czimmerer Z, Kampmann T, et al. Structural basis for the assembly of the SMRT/NCoR core transcriptional repression machinery. Nat Struct Mol Biol. 2011;18(2):177-84.
27. Cardamone MD, Krones A, Tanasa B, Taylor H, Ricci L, Ohgi KA, et al. A protective strategy against hyperinflammatory responses requiring the nontranscriptional actions of GPS2. Molecular cell. 2012;46(1):91-104.
28. Lee TH, Yi W, Griswold MD, Zhu F, Her C. Formation of hMSH4-hMSH5 heterocomplex is a prerequisite for subsequent GPS2 recruitment. DNA repair. 2006;5(1):32-42.
29. Sanyal S, Bavner A, Haroniti A, Nilsson LM, Lundasen T, Rehnmark S, et al. Involvement of corepressor complex subunit GPS2 in transcriptional pathways governing human bile acid biosynthesis. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(40):15665-70.
30. Holaska JM, Wilson KL. An emerin “proteome”: purification of distinct emerin-containing complexes from HeLa cells suggests molecular basis for diverse roles including gene regulation, mRNA splicing, signaling, mechanosensing, and nuclear architecture. Biochemistry. 2007;46(30):8897-908.
31. Diederichs S, Baumer N, Ji P, Metzelder SK, Idos GE, Cauvet T, et al. Identification of interaction partners and substrates of the cyclin A1-CDK2 complex. The Journal of biological chemistry. 2004;279(32):33727-41.
32. Peng YC, Kuo F, Breiding DE, Wang YF, Mansur CP, Androphy EJ. AMF1 (GPS2) modulates p53 transactivation. Mol Cell Biol. 2001;21(17):5913-24.

33. Zhang D, Harry GJ, Blackshear PJ, Zeldin DC. G-protein pathway suppressor 2 (GPS2) interacts with the regulatory factor X4 variant 3 (RFX4_v3) and functions as a transcriptional co-activator. The Journal of biological chemistry. 2008;283(13):8580-90.
34. te Velthuis AJ, Admiraal JF, Bagowski CP. Molecular evolution of the MAGUK family in metazoan genomes. BMC evolutionary biology. 2007;7:129.
35. de Mendoza A, Suga H, Ruiz-Trillo I. Evolution of the MAGUK protein gene family in premetazoan lineages. BMC Evol Biol. 2010;10:93.
36. Oliva C, Escobedo P, Astorga C, Molina C, Sierralta J. Role of the MAGUK protein family in synapse formation and function. Developmental neurobiology. 2012;72(1):57-72.
37. Zhu J, Shang Y, Xia C, Wang W, Wen W, Zhang M. Guanylate kinase domains of the MAGUK family scaffold proteins as specific phospho-protein-binding modules. EMBO J. 2011;30(24):4986-97.
38. Traweger A, Toepfer S, Wagner RN, Zweimueller-Mayer J, Gehwolf R, Lehner C, et al. Beyond cell-cell adhesion: Emerging roles of the tight junction scaffold ZO-2. Tissue barriers. 2013;1(2):e25039.
39. Gardoni F. MAGUK proteins: new targets for pharmacological intervention in the glutamatergic synapse. European journal of pharmacology. 2008;585(1):147-52.
40. Ma H, Cai H, Zhang Y, Wu J, Liu X, Zuo J, et al. Membrane palmitoylated protein 3 promotes hepatocellular carcinoma cell migration and invasion via up-regulating matrix metalloproteinase 1. Cancer Lett. 2014;344(1):74-81.
41. Roberts S, Delury C, Marsh E. The PDZ protein discs-large (DLG): the 'Jekyll and Hyde' of the epithelial polarity proteins. The FEBS journal. 2012;279(19):3549-58.
42. Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature. 2012;488(7409):106-10.
43. Zhang J, Yang X, Wang Z, Zhou H, Xie X, Shen Y, et al. Structure of an L27 domain heterotrimer from cell polarity complex Patj/Pals1/Mals2 reveals mutually independent L27 domain assembly mode. The Journal of biological chemistry. 2012;287(14):11132-40.
44. Bohl J, Brimer N, Lyons C, Vande Pol SB. The stardust family protein MPP7 forms a tripartite complex with LIN7 and DLG1 that regulates the stability and localization of DLG1 to cell junctions. The Journal of biological chemistry. 2007;282(13):9392-400.
45. Feng W, Long JF, Fan JS, Suetake T, Zhang M. The tetrameric L27 domain complex as an organization platform for supramolecular assemblies. Nature structural & molecular biology. 2004;11(5):475-80.
46. Gonzalez-Mariscal L, Bautista P, Lechuga S, Quiros M. ZO-2, a tight junction scaffold protein involved in the regulation of cell proliferation and apoptosis. Annals of the New York Academy of Sciences. 2012;1257:133-41.
47. Ojeh N, Pekovic V, Jahoda C, Maatta A. The MAGUK-family protein CASK is targeted to nuclei of the basal epidermis and controls keratinocyte proliferation. Journal of cell science. 2008;121(Pt 16):2705-17.
48. Egawa T, Albrecht B, Favier B, Sunshine MJ, Mirchandani K, O'Brien W, et al. Requirement for CARMA1 in antigen receptor-induced NF-kappa B activation and lymphocyte proliferation. Current biology : CB. 2003;13(14):1252-8.

49. Boormans JL, Korsten H, Ziel-van der Made AJ, van Leenders GJ, de Vos CV, Jenster G, et al. Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. International journal of cancer Journal international du cancer. 2013;133(2):335-45.
50 Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics. 2008 Aug 1;24(15):1707-14.
51. Eussen BH, van de Laar I, Douben H, van Kempen L, Hochstenbach R, De Man SA, et al. A familial inverted duplication 2q33–q34 identified and delineated by multiple cytogenetic techniques. European Journal of Medical Genetics. 2007;50(2):112-9.

Supplementary data

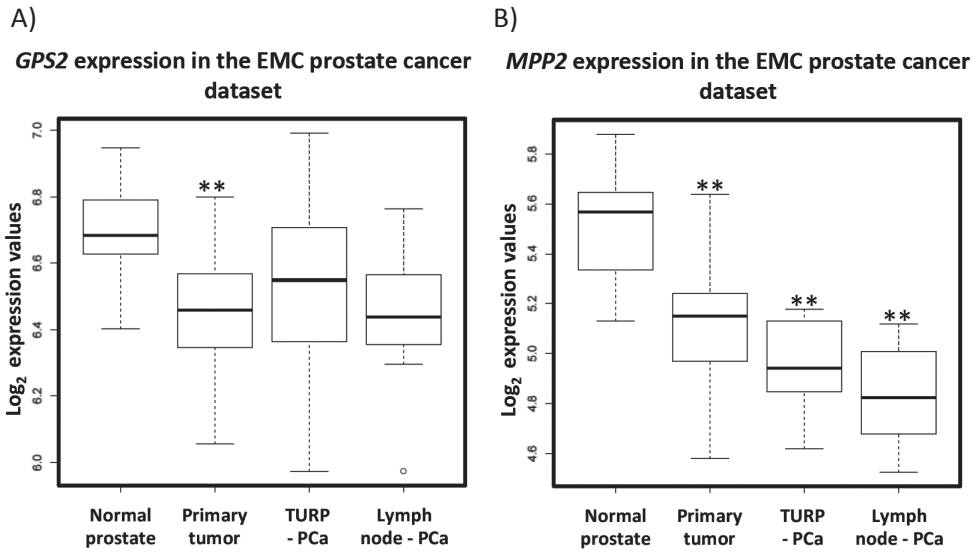


Figure S1: Expression levels of *GPS2* (A) and *MPP2* (B) in primary prostate cancer (56 samples), normal adjacent prostate tissue (12 samples), prostate cancer lymph node metastasis (12 samples) and transurethral resections (TURP, 10 samples). Wilcoxon testing was used to determine significant differences in expression levels between primary tumor and normal, lymph node metastasis and TURP. ** p-value ≤ 0.02.

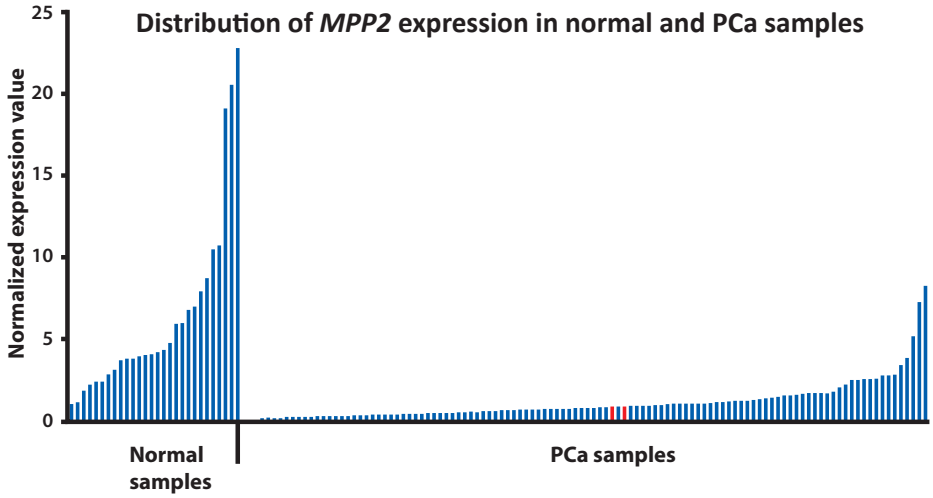


Figure S2: Expression levels of *MPP2* in normal prostate samples and in PCa samples. The expression levels of *MPP2* were measured by qPCR using a Taqman gene probe. The expression levels of *MPP2* measured by qPCR were normalized to the expression levels of *GAPDH* mRNA in each sample. In red, the expression levels of two LNCaP substrains (ATCC and Rotterdam) are highlighted.

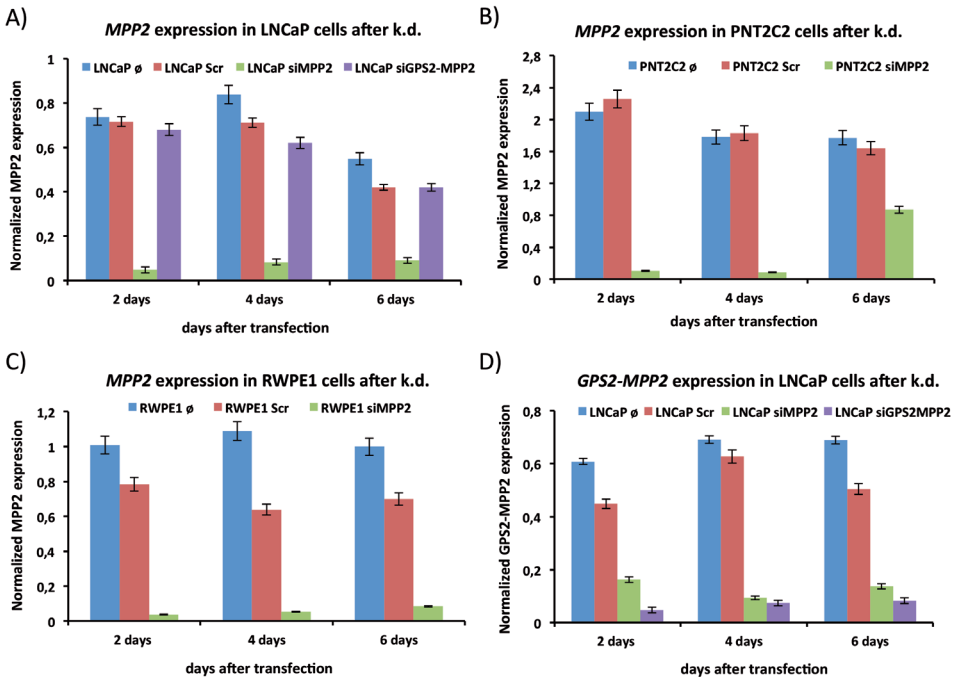


Figure S3: Taqman gene expression assay for *MPP2* and *GPS2-MPP2*. The expression level of *MPP2* was measured two, four and six days after siRNA transfection against *MPP2* and *GPS2-MPP2* in LNCaP (A) and against *MPP2* in PNT2C2 (B) and RWPE-1 (C) cell lines. In addition, the specific knock-down level of the *GPS2-MPP2* gene fusion was measured in the LNCaP cell line (D). All values were normalized to the expression of the *GAPDH* house keeping gene. The ø represents LNCaP cells growing under normal conditions and Scr the addition of a non-targeting small interfering RNA.

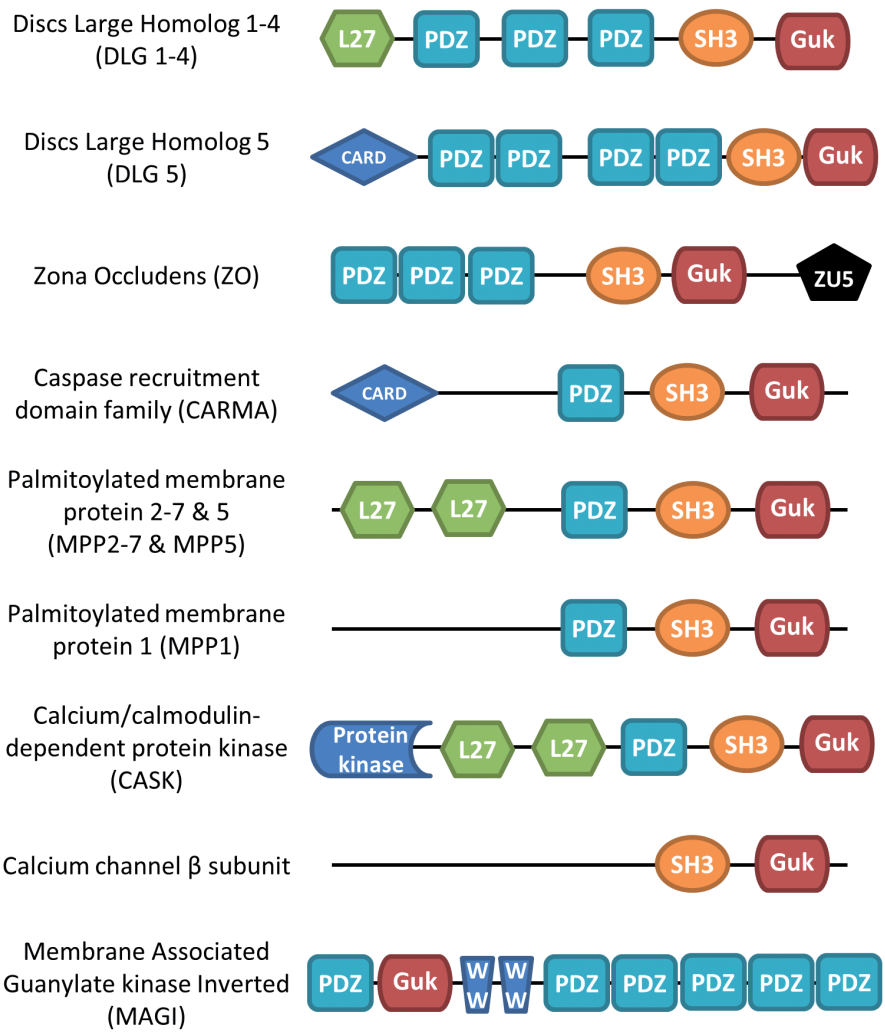


Figure S4: Scheme displaying the protein domain distribution in all the different classes of the MAGUK protein family. The L27 domain is found in the receptor targeting proteins Lin-2 and Lin-7 and is a specific protein-protein interaction module. PDZ domains are structural domains of 80 to 90 amino acids and usually bind to short C-terminal regions of other proteins. SH3 (SRC Homology 3) domains are found commonly in several intracellular or membrane-associated proteins. GuK (Guanylate kinase) domain. CARD (caspase activation and recruitment) domains are interaction motifs found in proteins involved typically in apoptotic processes and they mediate the formation of large protein complexes. ZU5 is a domain of 90 to 110 residues that in most cases contain a C-terminal death domain. WW is a short conserved region that binds proline-rich polypeptides.

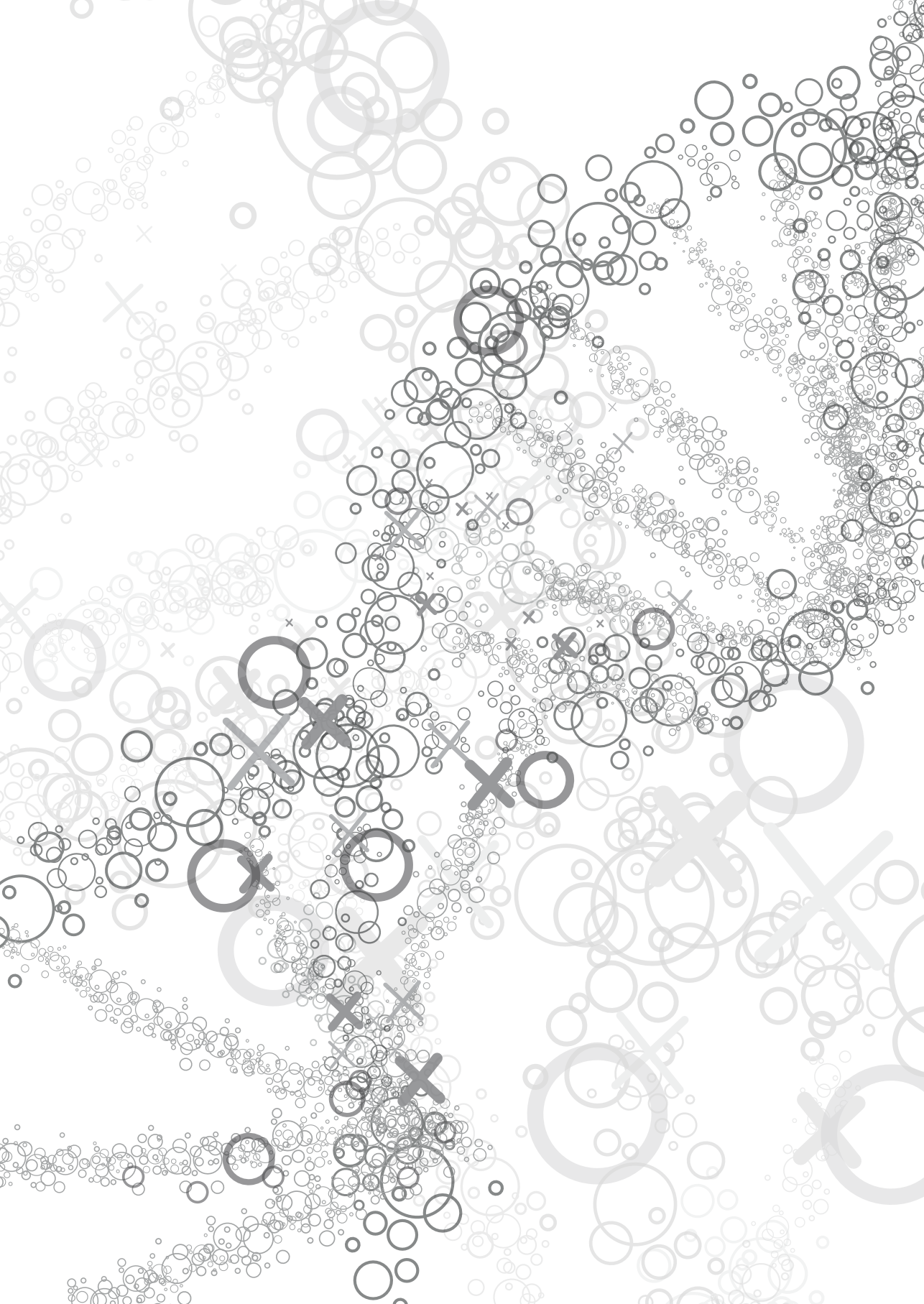
Supplementary Table 1, Supplementary Table 2 and Supplementary videos 1-4 will be made available through the journal's website upon publication.

Before publication they will be available through the following links:

Supplementary Table 1 -
https://www.dropbox.com/sh/qe2lx9obl3q4mel/AADhlj0J-Ox-B5XWUwz7c2caa?dl=0

Supplementary Table 2 -
https://www.dropbox.com/sh/qe2lx9obl3q4mel/AADhlj0J-Ox-B5XWUwz7c2caa?dl=0

Supplementary videos 1-4 -
https://www.dropbox.com/sh/jyih0ntzfc64zc0/AAAPTD11w1dUVNCuEDSwMbNZa?dl=0



Chapter 3

Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer

I Teles Alves[1,2]
T Hartjes[2]
E McClellan[3,4]
S Hiltemann[1]
R Bottcher[1]
N Dits[1]
MR Temanni[5]
B Janssen[6]
W van Workum[6]
P van der Spek[3]
A Stubbs[3]
A de Klein[7]
B Eussen[7]
J Trapman[2] and G Jenster[1]

Departments of Urology[1], Pathology[2], Bioinformatics[3] and Clinical Genetics[7], Erasmus MC, Rotterdam, The Netherlands

Department of Mathematical and Computer Sciences[4], Metropolitan State University of Denver, Colorado, USA;

CMNS-Institute for Advanced Computer Studies[5], University of Maryland, College Park, MD, USA;

ServiceXS B.V.[6], Leiden, The Netherlands;

Oncogene 2015; 34(5):568-77

Abstract

Gene fusions, mainly between TMPRSS2 and ERG, are frequent early genomic rearrangements in prostate cancer (PCa). In order to discover novel genomic fusion events, we applied whole-genome paired-end sequencing to identify structural alterations present in a primary PCa patient (G089) and in a PCa cell line (PC346C). Overall, we identified over 3800 genomic rearrangements in each of the two samples as compared with the reference genome. Correcting these structural variations for polymorphisms using whole-genome sequences of 46 normal samples, the numbers of cancer-related rearrangements were 674 and 387 for G089 andPC346C, respectively. From these, 192 in G089 and 106 in PC346C affected gene structures. Exclusion of small intronic deletions left33 intergenic breaks in G089 and 14 in PC346C. Out of these, 12 and 9 reassembled genes with the same orientation, capable of generating a feasible fusion transcript. Using PCR we validated all the reliable predicted gene fusions. Two gene fusions were in-frame: MPP5–FAM71D in PC346C and ARHGEF3–C8ORF38 in G089. Downregulation of FAM71D and MPP5–FAM71D transcripts inPC346C cells decreased proliferation; however, no effect was observed in the RWPE-1-immortalized normal prostate epithelial cells. Together, our data showed that gene rearrangements frequently occur in PCa genomes but result in a limited number of fusion transcripts. Most of these fusion transcripts do not encode in-frame fusion proteins. The unique in-frame MPP5–FAM71D fusion product is important for proliferation of PC346C cells.

Introduction

Prostate cancer (PCa) is one of the most frequently diagnosed cancers and a major cause of death in men in countries with a western lifestyle (1). Throughout the past two decades, several genetic events have been revealed that are important in development and progression of PCa (2). The predominant genetic abnormalities identified so far include the forming of ETS-fusion genes (3), loss of phosphatase and tensin homolog (PTEN) tumour suppressor gene (4), amplification of AR and amplification of the MYC oncogene (5). Most of these studies used comparative genomic hybridization/SNP (single nucleotide length polymorphism) arrays and performed genome-wide copy number variation analysis as a start to identify specific genetic alterations (6, 7). The use of gene expression arrays was important in the detection of genes with aberrant expression patterns (8). The integration of these two different sets of data, copy number variation and gene expression reinforced and expanded this panel of genetic alterations (9, 10).

Next-generation sequencing (NGS) techniques emerging over the last few years proved to be a major breakthrough in documenting novel genetic changes resulting in a better understanding of cancer cell biology (11, 12). Both RNA and DNA can be used as templates for NGS methods and it is possible to analyse either paired-end, mate pair, as short or long sequence reads depending on the platform applied (13, 14). An exome-sequencing approach uses only 1–2% of the genomic sequences as a template through capture and enrichment before the sequencing process (15, 16). This allows for higher sequence coverage at the expense of a few disadvantages such as the uneven capture efficiency and the absence of unknown or yet to be annotated exons (17). Instead of a focused exome-sequencing approach, more challenging whole-genome sequencing provides the most complete view of genomic changes (18). A range of sequencing technologies is now available and more are becoming available soon that provide different approaches for library construction, clone separation and amplification, and nucleotide detection (19–22). The technology utilized in this study from Complete Genomics Inc. (Mountain View, CA, USA) makes use of array technology to separate amplified DNA template organized into single-strand coils, known as DNA nanoballs. Nucleotide read-out is based on a ligation protocol (23).

In PCa genomics the use of NGS has allowed significant progress in cataloguing systematically all the DNA changes present in cancer (24). The use of RNA-seq has produced major insight into the identification and expression of novel long noncoding RNAs and novel gene fusions in PCa (25, 26). Exome sequencing of PCa samples was utilized for the detection of novel small mutations (27–30). In the same way, a mutational landscape of patients with castration-resistant PCa was described by Tomlins et al. (3) in 2012 (31). In addition, deep RNA sequencing is also used to identify transcription-induced chimeras associated with human prostate adenocarcinoma such as the novel TMEM79-SMG5 (32, 33). However, a full overview of all structural variations (SVs) and mutations can only be provided by whole-genome sequencing. Many studies so far have applied NGS to obtain a more thorough perceptive of already known genetic alterations

(34). The mechanism involving PTEN loss has been correlated with novel genetic alterations in genes located in the vicinity of PTEN (24). Moreover, the expression of constitutively active androgen receptor splice variants in castration resistance PCa has been described (35). A few studies have been published making use of whole or focused genome NGS technologies to discover and describe new genetic alterations in PCa. The study by Berger et al. (24) in 2011 provided a comprehensive approach towards both known and newly identified mutations and genomic rearrangements. Other studies have focused on either a particular genomic region such as the 10q11.2 97-kb region comprising the MSMB gene or in the analysis of all alterations present in a particular PCa sample (36, 37). This was the case for the new type of prostate adenocarcinoma identified that has a hybrid phenotype of both luminal and neuroendocrine cells (37, 38).

In this study, we applied whole-genome paired-end sequencing in two PCa samples. Analysis of the PC346C PCa cell line and the G089 primary PCa patient sample, which were selected based on the absence of ETS-fusion genes (39), generated a wide panel of new SV events. We focused on the detection of SVs to unravel the events leading to fusion genes. The potential fusion genes were validated and the genes involved were checked for copy number alteration in other PCa samples. The in-frame fusion found in PC346C was assessed for possible functional implications.

Results

SVs detected by NGS

Whole-genome paired-end sequencing (Complete Genomics Inc.) of two PCa samples, the G089 primary tumor and the PC346C cell line, generated data with an average coverage of 61x and 67x, respectively (Supplementary Table S1, Supplementary Figure S1). The fully aligned genome fraction for both samples was above 93% (Supplementary Figure S2). Data showing reads mapping at a different position/orientation than expected were used to perform de novo assembly and generate a list of all SVs in the two DNAs. In the PC346C sample, a total of 3898 SVs were detected (both interchromosomal and intrachromosomal) as compared with the reference genome. For the G089 sample the total of SVs was 3837 (Table 1). As normal DNA from these two PCa samples was not available, the data were curated for polymorphisms using a data set derived from 46 normal DNA samples (Supplementary Table S2). This left us with 387 candidate cancer-related events in PC346C and 674 in G089 (Table 1).

Overall, 90.1% and 82.4% of all SVs detected in PC346C and G089 were also present in any of the 46 normal controls. Of the cancer related events the far majority did not occur in a gene intron or exon: only 2.7% and 5.0% of SVs were inside genes for both junctions in PC346C and G089, respectively (Figures 1a and b).

Table 1: Structural variations (SVs) identified by next-generation sequencing

Type of SV	Prostate cancer samples (before correction by 46 normal samples					
	Total number		Gene only on one side		Genes on both sides	
	PC346C	G089	PC346C	G089	PC346C	G089
Interchromosomal	1228	1283	249	231	36	44
Intrachromosomal	2670	2554	48	67	796	794
Total	3898	3837	297	298	832	838
	46 normal samples (average per one sample)					
	Total number		Gene only on one side		Genes on both sides	
Interchromosomal	120		21		4	
Intrachromosomal	301		5		94	
Total	421		26		98	
	Prostate cancer samples (after correction by 46 normal samples)					
	Total number		Gene only on one side		Genes on both sides	
	PC346C	G089	PC346C	G089	PC346C	G089
Interchromosomal	86	141	25	52	5	18
Intrachromosomal	301	533	17	44	101	174
Total	387	674	42	96	106	192

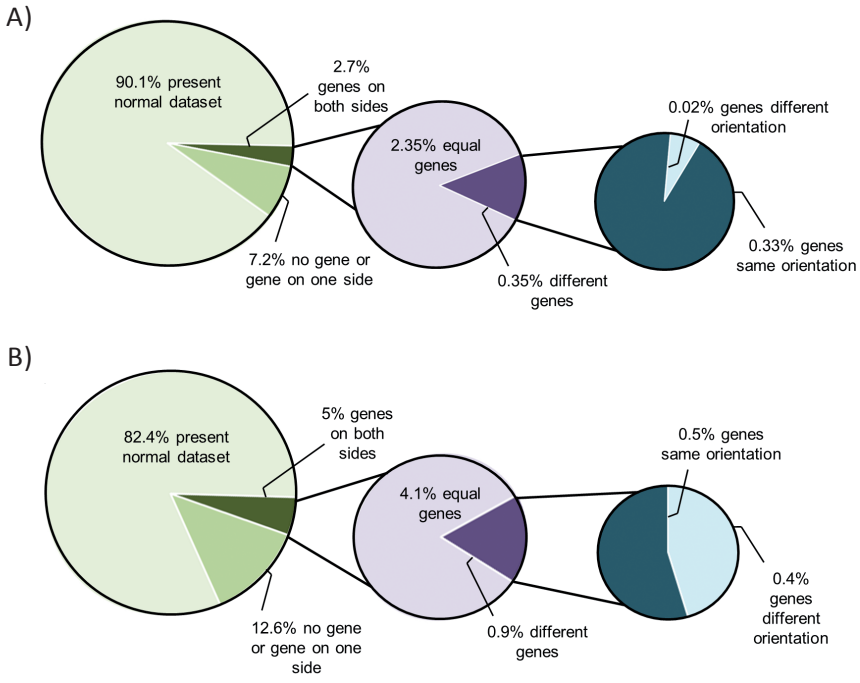


Figure 1: Overview of genomic abnormalities identified by NGS. Pie chart illustrating the genomic breaks detected by NGS in PC346C (A) and G089 (B). The percentages represent the presence of junctions in any of the 46 normal samples (light green) and their occurrence within genes (green and dark green). Next, we classify according to the overlap of junctions with different genes (light and dark purple) and the gene orientation (light and dark blue).

Additionally, we often found the same gene IDs on both sides of the junctions, which corresponded to (small) intragenic rearrangements, such as deletions, duplications and inversions. We detected 14 SV events in PC346C and 33 in G089 where each side of the junction occurred within a different gene (Figure 2) (40).

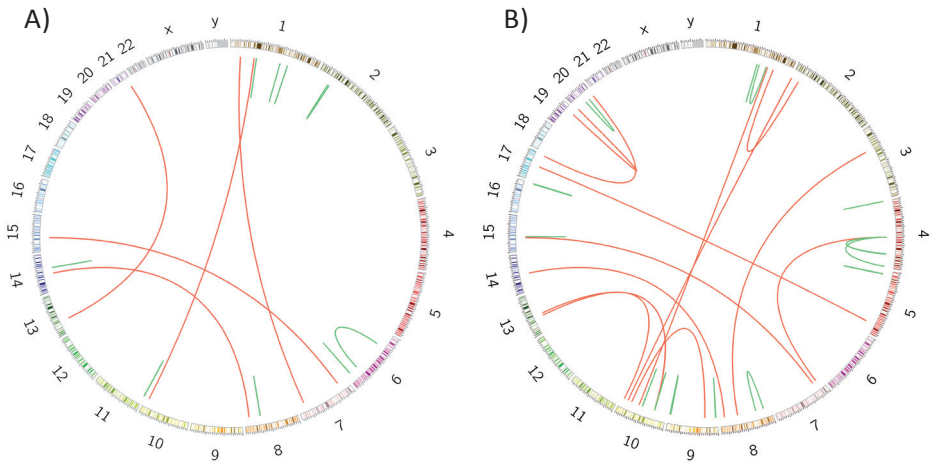


Figure 2: Graphical representation of gene rearrangements in PC346C and G089. Circos plot of the PC346C (A) and G089 (B) samples. The outer ring depicts the chromosomes. The 14 and 33 gene rearrangements found in PC346Cand G089, respectively, are shown in red (interchromosomal) and green (intrachromosomal) lines.

From these 14 and 33 SVs, we excluded the events in which the orientation of the genes was head-to-head or tail-to-tail to select for gene fusion events that could generate fusion transcripts. In the end, nine (PC346C) and 11 (G089) SVs fusing two different genes in the same orientation were identified (Table 2; Supplementary Tables S3–S5).

Validation of fusion genes present in PC346C and G089

For each possible gene fusion event in PC346C and G089, we checked all the mapped reads, both the normal and the mispaired reads (Supplementary Figure S3). We observed that a number of fusion events were not supported by a convincing number of discordant mate pairs (Table 2). To evaluate whether the criterion of three or more mispaired reads was justified, all 9 + 11 fusion events were validated using (RT)–PCR both at the DNA and RNA levels (Supplementary Figures S4 and S5, Supplementary Tables S6–S9). Most of the fusion events with low mispaired read counts (≤ 5) could not be confirmed, showing that the cut off number of three discordant mate pairs was set low.

For most gene fusions, the event was confirmed by sequencing on the RNA and/or DNA level. Out of these six PC346C fusion events, all were validated on the DNA level with

Table 2: List of all candidate gene fusions detected by NGS and their validation on the RNA and DNA levels

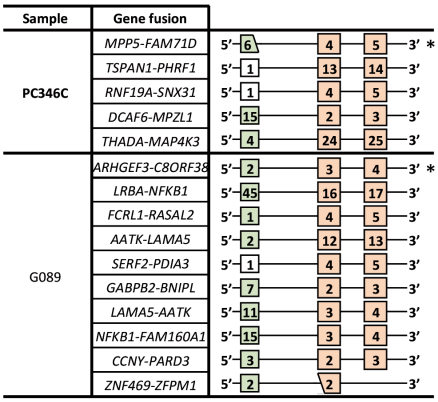
Sample	5' Donor gene	3' Acceptor gene	5' Chr	3' Chr	Validated cDNA	Validated DNA	In frame	# of Discordant mate pairs
PC346C	ATP6V0A4	PHACTR4	chr7	chr1	No	No	-	5
	GTF2H5	SYNJ2	chr6	chr6	No	No	-	3
	MPP5	FAM71D	chr14	chr14	Yes	Yes	Yes	43
	TSPAN4	PHRF1	chr11	chr11	Yes	Yes	No	62
	RNF19A	SNX31	chr8	chr8	Yes	Yes	No	16
	DCAF6	MPZL1	chr1	chr1	Yes	Yes	No	29
	PTCRA	ENPP3	chr6	chr6	No	No	-	3
	THADA	MAP4K3	chr2	chr2	Yes	Yes	No	55
	ITGA10	RBM8A	chr1	chr1	No	Yes	-	24
G089	ARHGEF3	C8orf38	chr3	chr8	Yes	Yes	Yes	39
	LRBA	NFKB1	chr4	chr4	Yes	Yes	No	27
	FCRL1	RASAL2	chr1	chr1	No	Yes	-	82
	AATK	LAMA5	chr17	chr20	No	Yes	No	5
	SERF2	PDIA3	chr15	chr15	Yes	No	No	8
	GABPB2	BNIPL	chr1	chr1	Yes	Yes	No	15
	LAMA5	AATK	chr20	chr17	Yes	Yes	No	41
	NFKB1	FAM160A1	chr4	chr4	No	Yes	-	58
	CCNY	PARD3	chr10	chr10	No	Yes	-	17
	OBP2A	OBP2B	chr9	chr9	Yes	No	Yes	3
	ZFPM1	ZNF469	chr16	chr16	No	Yes	-	20

Abbreviations: cDNA, complementary DNA; Chr, chromosome; NGS, next-generation sequencing

only one that could not be verified at the transcript level (Supplementary Tables S10 and S11). The fusion in question, *ITGA10–RBM8A*, comprises the downstream part of the 30-untranslated repeat sequence of *RBM8A* so it was not expected to be present within the fusion transcript. In G089, out of the nine reliable fusion events eight were validated at the DNA level and five at the RNA level. The four fusions that were validated only at the DNA level were assessed for the expression levels of the involving genes using Affymetrix Exon-array analysis (Affymetrix, Santa Clara, CA, USA) of the G089 PCa sample. In general, we found a low expression pattern of these fusion genes. The *OBP2A–OBP2B* fusion, comprising only three mispaired reads, was verified at the RNA level and involved two homologous genes. The *AATK–LAMA5* fusion was confirmed at the DNA level, although the number of discordant mate pairs was only five (Table 2).

Next, we assessed which fusion events generated an in-frame fusion protein. On the basis of the sequence of the fusion transcript we determined that *MPP5–FAM71D* in PC346C and *ARHGEF3–C8ORF38* in G089 were in-frame and most likely produced a fusion protein (Table 2). The final set of validated gene fusions (Figures 3a and b) was assessed for discordant mate pairs on both sides of the break as well as visible copy number-related breakpoints using the Illumina 1M SNP array (Illumina). The presence of discordant mate pair reads on both sides of the break corresponds to distinct genomic

A)



B)

	Discordant mate pairs on both sides of the break		Break visible at the SNP level	
Gene fusions	5′ donor	3′ acceptor	5′ donor	3′ acceptor
MPP5-FAM71D	no	no	no	no
TSPAN4-PHRF1	no	no	yes	yes
RNF19A-SNX31	no	no	no	no
DCAF6-MPZL1	no	no	yes	yes
THADA-MAP4K3	no	no	yes	yes
ARHGEF3-C8ORF38	no	yes (chr8:96506639)	yes	no
LRBA-NFKB1	yes (chr7:11346366)	yes (*FAM160A1*)	no	no
FCRL1-RASAL2	no	no	no	yes
AATK-LAMA5	no	yes (*LAMA5*)	no	no
SERF2-PDIA3	no	no	no	no
GABPB2-BNIPL	no	no	no	no
LAMA5-AATK	yes (*AATK*)	no	no	no
NFKB1-FAM160A1	yes (*LRBA*)	yes (chr7:12215681)	no	no
CCNY-PARD3	yes (*WDFY2*)	no	yes	yes
ZNF469-ZFPM1	no	no	yes	yes

Figure 3: Validated gene fusions detected by NGS. (A) DNA representation of the gene fusions validated in PC346C and in G089. The oblique line in *MPP5* and ZFPM1 denotes the break occurs within the exon. The white rectangles correspond to non-coding exons. Light green and orange squares correspond to the acceptor and donor genes, respectively. All the fusions represented are validated at the RNA and/or DNA level. The * denotes in-frame transcripts. (B) All the gene fusions were checked for the presence of discordant mate pairs on both sides of the break as well as for breaks visible at the SNP level.

locations that were joined together. On the other hand, a break with discordant mate pairs on only one side will, most likely, correspond to a break with loss of DNA on the other side. Regarding our final set of 15 fusions we observe mainly the presence of discordant mate pairs on one side of the break (Figure 3b) corresponding to the candidate gene fusion we initially detected.

The remaining fusions had discordant mate pairs on both sides, meaning both sides of the break fused to different genomic locations. That was the case for *AATK–LAMA5* with the reciprocal *LAMA5–AATK* gene fusion detected, *NFKB* that fuses to both *LRBA* and *FAM160A1* and *CCNY*, which fuses to *PARD3* and *WDFY2*. The remaining *C8ORF38, LRBA* and *FAM160A1* gene fragments break to *ARHGEF3* and *NFKB*, respectively, as well as to additional intergenic genomic locations.

All the genes of the validated fusion events were checked for breakpoints in a 66 SNP array data set of 11 PCa xenografts, eight PCa cell lines and 47 PCa patient samples with six matching normal controls. We observed that a few genes had breakpoints at the SNP level in other PCa samples providing us with an impression of the frequency of the specific breaks (Supplementary Table S12, Supplementary Figure S6). We checked for the presence of the same fusion transcript (*THADA–MAP4K3, ARHGEF3–C8ORF38 and LRBA–NFKB1*) in the samples with copy number variations (CNVs), but never observed a positive RT–PCR using the primer pairs that proved the fusion transcripts in G089and PC346C (data not shown).

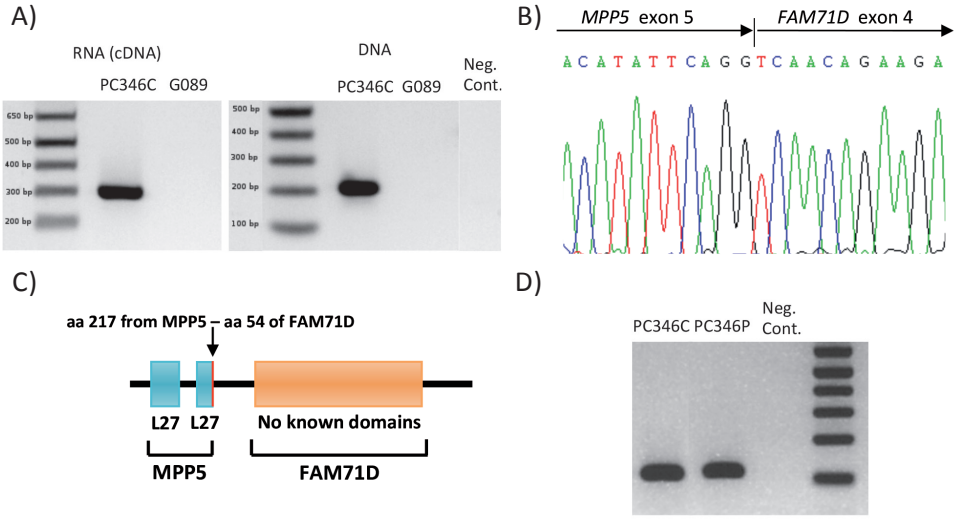


Figure 4: Discovery of the novel *MPP5-FAM71D* fusion in PC346C. (A) Validation of *MPP5-FAM71D* fusion transcript in RNA (cDNA) and DNA from PC346C by PCR amplification and electrophoresis. The fusion is only present in PC346C and not in G089. The third lane corresponds to the negative control (no cDNA added). (B) Validation of the *MPP5-FAM71D* fusion breakpoint using Sanger sequencing in cDNA from PC346C. (C) Scheme of the *MPP5-FAM71D* fusion protein. The red line corresponds to the disruption of the second L27 domain which retains the first 40 aa from the original 57 aa. *FAM71D* has no known structural domains. (D) The presence of the *MPP5-FAM71D* fusion (cDNA) was tested in the PC346P patient-derived xenograft from which the PC346C cell line originated. The presence of the band in both PC346C and PC346P shows that this fusion was already present in the xenograft.

MPP5–FAM71D fusion has functional implications in PC346Ccell line

The two in-frame gene fusions were assessed for the expression level of the 30 acceptor genes in the exon-array data of G089 and PC346C. We observed a clear upregulation of *FAM71D* expression from exon 4 onwards in the PC346C sample (Supplementary Figure S7) compared with normal controls and with G089. Regarding *C8ORF38* we did not observe differential expression in G089 as compared with normal controls and other PCa samples (Supplementary Figure S8). *MPP5* and *ARHGEF3* do not show differential expression in PCa samples or in PC346C and G089.

The *MPP5–FAM71D* fusion was confirmed both at the RNA and DNA levels (Figure 4a) and the breakpoint was sequenced (Figure 4b). Forty-three discordant mate pairs were detected for *MPP5* and *FAM71D* (Table 2). The two genes are located on chromosome 14, ~52 kb apart. In the SNP array data we did not detect any copy number alteration (deletion or gain) in these two genes. The fusion transcript predicted to code for a fusion protein of

aa 1–217 of *MPP5* (674 aa full-length protein) and aa 54–422 of *FAM71D* (Figure 4c). The expression of *MPP5* is not known to be prostate-specific or androgen-regulated. Obviously, the expression of the 30-sequence of *FAM71D* was highly upregulated in PC346C, now being under control of the *MPP5* promoter region. The PC346P patient-derived xenograft from which the PC346C cell line originated also shows high expression of the fusion gene (Figure 4d).

In order to assess the frequency of the *MPP5–FAM71D* fusion and *FAM71D* overexpression, we performed TaqMan gene expression assays on 201 PCa samples. None of these were positive for the *MPP5–FAM71D* fusion transcript observed in PC346C. In order to extend the analyses to other genomic events that would result in the *FAM71D* overexpression, we found 10 samples to have a relative high expression of *FAM71D* as compared with PC346C, normal prostate and the bulk of cancer samples (Supplementary Figure S9).

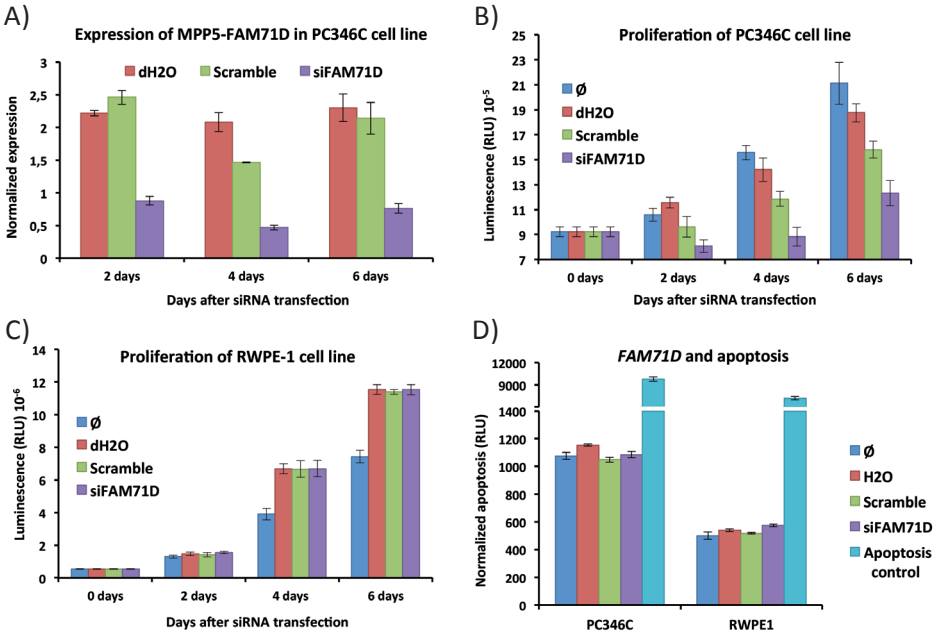


Figure 5: *FAM71D* promotes cell proliferation in the PC346C cell line. (A) Knockdown of *FAM71D* in the PC346C cell line. The levels of *FAM71D* expression were still downregulated at day 6 after the knockdown. Data were normalized to *GAPDH* and represented as fold-change. (B) Cell proliferation assays for PC346C cells. The ø represents PC346C cells growing under normal conditions, dH2O the addiction of transfection reagent, scramble the addition of a non-targeting siRNA and siFAM71D the knockdown of the *FAM71D* transcript. An asterisk (*) indicates $P \leq 0.02$ and (**) $P \leq 0.002$ by a two-tailed Student's t-test. (C) Cell proliferation assays for RWPE1. Same controls as used for the PC346C cells. (D) Apoptosis assay through detection of Caspase 3/7 activity corrected for cell viability. This assay was performed for both PC346C and RWPE-1 cell lines. The same controls of the proliferation assays were used along with an apoptosis control (drug-treated cells for 24 h).

Next, we addressed the functional relevance of *FAM71D* overexpression in the PC346C cell line by the knockdown of its expression. The small interfering RNA SMARTpool of *FAM71D* successfully caused downregulation in both PC346C and the normal epithelial prostate RWPE-1 cell line (Figure 5a). In the case of PC346C cells we observed a considerable decrease in proliferation capacity at all time points tested (2, 4 and 6 days after the knockdown). This decrease was significant when compared with all the transfection controls and the normal PC346C growth (Figure 5b). In contrast, RWPE-1 cells showed a slight increase in proliferation after *FAM71D* knockdown, which was not significant when compared with the transfection controls (Figure 5c).

In order to address whether there was induction of apoptosis by *FAM71D* knockdown, we measured the caspase 3/7 activity for both transfected PC346C and RWPE-1. We did not detect an increase in apoptosis in PC346C or RWPE-1 due to the knockdown of *FAM71D* as compared with the transfection controls (Figure 5d). The cell growth effect of overexpression of *FAM71D* in PC346C is most likely the result of increased proliferation and not inhibition of apoptosis. The selective decrease in proliferation only in PC346C indicates a functional beneficial role for *MPP5–FAM71D* in these cancer cells.

Discussion

Correction for most SV polymorphisms using a control data set of 46 normal samples

In the present study, we report the SVs detected by whole genome paired-end sequencing of two PCa samples, the primary tumor G089 and the cell line PC346C. As described above, we detected 3898 and 3837 SVs in PC346C and G089 as compared with the HG18 reference genome. As most of these SVs are polymorphisms, one needs to separate these common variants from the cancer-related events. Optimally, the sequence of a normal DNA sample from the cancer patient will do so, but the normal control DNA of the G089 patient sample and PC346C cell line is not available. This is not an uncommon problem since for many cell lines, xenografts and old cancer sample repositories, a normal blood or tissue sample has not been stored. In order to remove common polymorphism, one can make use of databases of structural variants that are far from complete but are becoming more comprehensive rapidly (41). We decided to use a control dataset of 46 normal DNA samples that were sequenced on the same Complete Genomics platform and analyzed using the identical pipeline. This resulted in the exclusion of 3511 and 3163 SVs from the initial list of SVs detected in PC346C and G089, respectively. As expected, the majority of the SVs detected initially were normal occurring variation across individuals. In the end, we anticipate that most of these common variants will not correlate with cancer, although we cannot exclude their contribution to disease susceptibility (42, 43).

The list of gene-to-gene rearrangements in the 46 normal samples was used to estimate a false discovery rate of cancer related gene fusions. We observed that 28% of all

the 'common' gene-to-gene SVs were present in only one of the 46 normal samples (Supplementary Table S6). Such a high proportion means that many gene-to-gene SVs are rare and therefore we can assume that of our cancer-related gene fusions described, some are rare polymorphisms.

Overall, our study showed that, in the absence of normal control DNA, it is still possible to eliminate the majority of the normal polymorphisms. Correction for rare polymorphisms will improve with the use of larger sets of normal control samples particularly with samples of the same ethnic origin.

G089 and PC346C PCa samples do not show positive selection for breaks inside genes or in-frame gene fusion transcripts

The detection of novel gene fusions in G089 and PC346C aimed at investigating the relevance of these events in the initiation and/or progression of PCa. As gene rearrangements often affect the genes involved and are therefore under selection pressure, one might expect these events to be more common than random nonfunctional intergenic SVs. In order to determine whether there is a bias towards breaks inside genes, we checked whether such occurrences are more prevalent than the 38.5%—the approximate percentage of our genome that consists of genes (introns and exons) (44). Considering that each SV consists of two breakpoints, out of the 774 breaks in PC346C (387 SVs, Table 1), 42 and 212 (106 SVs with genes on both sides of the SV) are inside a gene, which is 32.8% of all breaks observed. For G089, in which out of the 1348 breaks, 96 and 384 are inside a gene representing 35.6% of all the breaks observed. This shows there is no bias towards breaks inside genes in our two PCa samples. Apparently, the occurrence of DNA breaks is an arbitrary process and is not selective for genomic coding regions.

Most of the gene rearrangements present in PC346C and G089 corresponded to small deletions, insertions or inversions within the same gene. After the exclusion of these SVs, we were left with a total of 20 feasible gene fusions from which 17 were successfully validated. The remaining three fusions had a very low number of discordant mate pairs and are likely false positives.

In order to verify the presence of in-frame gene fusions, we sequenced the breakpoint of the fusion transcripts. We observed that three gene fusions generated an in-frame fusion transcript, which is less than expected by random exon splicing, and conclude that there is no bias towards in-frame gene fusions in PC346C and G089. We expect that many of the gene fusions we have identified will be passenger events and not genetic drivers of PCa. However, it is certainly possible that some of the fusion genes that are inactivated by the rearrangement have tumor suppressor activity and their knockout does contribute to tumor progression. We checked whether additional mutations were identified in the fusion genes and found that with the exception of a few (*THADA, C8ORF38, LRBA* and *LAMA5*) all represent novel genes with potential implications in cancer biology.

Integration of NGS data with SNP and gene expression arrays is a powerful approach to determine the most relevant SVs in cancer samples

In our study, we used SNP and Exon-array data samples in conjunction with the NGS data to find support for the mechanism and relevance of the rearrangement using the copy number variation and transcript expression changes.

Regarding the unique gene partners identified in our study, we used the gene expression data to exclude genes that are not expressed in either PC346C or G089. This was particularly important to determine the expression change of fusion gene acceptors. In relation to the in-frame gene fusions, we observed that *FAM71D* is clearly upregulated in PC346C as compared with other normal and PCa samples, whereas *C8ORF38* is not differentially expressed in G089 as compared with the other samples (Supplementary Figures S7 and S8). Owing to the fusion, *FAM71D* is regulated and expressed in a higher manner by the *MPP5* promoter, and this outlier expression is a good indication that this gene fusion has a functional role in PC346C (Supplementary Figure S10).

Our final set of gene fusions comprised several candidate genes, many of which were not yet identified as cancer-related. The genes *LRBA, KNF19A, PARD3* and *THADA* were reported previously to be involved in other genomic events in PCa and thyroid adenocarcinoma (*THADA*) (24, 32, 33, 45). *LRBA* was found to be rearranged with regions downstream of the gene *FSTL5*, whereas *KNF19A* was not directly affected since the breakpoint was 3 kb upstream of the coding region (24). As for *PARD3* it was found to be fused to *ARHGAP10* and *THADA* to noncoding regions in chr3p25 and chr7p15 (32, 33) SNP microarray data allowed us to check whether any of the respective gene partners of our list of fusion genes had additional breaks visible at the SNP level in an additional 64 PCa samples (Supplementary Figure S11). *THADA, MAP4K3, ARHGEF3* and *LRBA* were found to have copy number breaks in other PCa samples, mainly PCa xenografts, cell lines and late stage tumors (Supplementary Table S12). The same fusions (*THADA–MAP4K3, ARHGEF3–C8ORF38* and *LRBA–NFKB1*) were not present in these samples with the CNV, indicating that *THADA, MAP4K3, ARHGEF3* and *LRBA* rearrange to diverse genomic locations, potentially different fusion partners. The CNV check using SNP array data will underestimate the number of samples with breaks inside these genes, as rearrangements without loss of DNA and small deletion or amplifications cannot be identified using this technology. Recurrent breaks strongly suggest a role for these genes in PCa progression, although further studies are needed to explore the exact frequency and functional consequences.

The integration of NGS and SNP microarray data generated a few inconsistent observations. We expected the lack of discordant mate pair reads on both sides of the break to indicate loss or gain of genetic material and therefore to be detected as a copy number change in the SNP array. This was the case for eight of our gene fusion partners, in which only one side of the break revealed discordant mate pairs without CNV in the SNP array data. This discrepancy can be due to the lower resolution of the SNP

array analyses, which has difficulties in detecting small deletions and amplifications. In the future, small-window copy number estimate from high coverage NGS data might resolve the differences observed in our study.

MPP5–FAM71D is related to the proliferation capacity of PC346C

The *MPP5–FAM71D* fusion, identified in the PC346C cell line, was silenced by the knockdown of *FAM71D*. This resulted in decreased proliferation capacity of PC346C cells but had no effect on RWPE-1-immortalized normal epithelial cells. This indicates that *MPP5–FAM71D* has a functional role in the growth of PC346C, although the precise mechanism by which this fusion exerts its function is not clear. *FAM71D* is located on chromosome 14 in the vicinity of *MPP5* (46) and its function is largely unknown. Recent studies in African trypanosomes showed Fam71 to function as a calcium-binding protein through its EF-hand calcium-binding domain but nothing is known on the mammalian ortholog (47).

Although we could not detect the *MPP5–FAM71D* fusion in any other of the 201 PCa samples analyzed, we did find relative higher expression of *FAM71D* in 10 of these cancer samples. The basis of this higher expression is unclear and could be due to fusion events or transcriptional upregulation. We conclude that *FAM71D* overexpression is not a frequent event in PCa and that the fusion to *MPP5* as observed in PC346C is patient-specific and extremely rare.

Nevertheless, we cannot exclude an influence of *MPP5*, which is disrupted by the fusion event. *MPP5* is a member of the membrane-associated guanylate kinase family and contributes to the establishment of cell polarity in mammals, which is crucial for tissue organization and whose loss is a hallmark of cancer (48–50). So far, the loss of *MPP5* has led to defects in polarity and assembly of tight junctions (51) as well as ineffective delivery of E-cadherin to the cell surface (52). The importance of *MPP5* in several cellular functions can therefore have a role in this outcome. We can infer that the fusion will cause a disruption of normal *MPP5*, which is crucial for cell polarity. However, one copy of *MPP5* is expressed and intact (no rearrangement or mutation), which excludes a complete *MPP5* silencing mechanism. Further functional studies should be performed as to determine through which mechanism this fusion is directly affecting cell proliferation.

Materials and Methods

Samples and whole-genome sequencing data analysis

The DNA of two PCa samples, the PC346C PCa cell line (53) and the G089 PCa patient were sequenced by Complete Genomics (Complete Genomics Inc.). For microarray analyses and whole-genome sequencing, freshly frozen clinical samples were obtained from the tissue bank of the Erasmus University Medical Center. Collection of patient samples has been performed according to national legislation concerning ethical requirements. Use of these samples has been approved by the Erasmus MC Medical

Ethics Committee according to the Medical Research Involving Human Subjects Act (MEC-2004-261). NCBI build 36 (hg18) was used as a reference genome during the mapping and data analysis process. Rearrangements were identified from discordant paired sequence reads (discordant mate pairs), which were mapping to different chromosomes (translocations), different positions on the same chromosome 4400 bp apart (deletions, inversions and duplications) or in unexpected orientations (small inversions and tandem duplications). The events are selected based on a minimum threshold of three discordant mate pairs having the same event. However, potential rearrangements were removed if there were any supporting discordant pairs for the same event in a panel of additional normal genomes that had already been sequenced by Complete Genomics (23) (Supplementary Methods).

Illumina 1M SNP Array analysis

SNP array data from 47 PCa patients, 11 PCa xenografts, eight PCa cell lines and six non cancer patients were collected. Genotyping was performed using the Infinium Illumina Human 1M probe BeadChip containing 1'072'820 markers, among which 206 665 are in reported CNV regions. The experiments were performed according to the manufacturer's protocol (Illumina) by an accredited service provider (ServiceXS). Data analysis was performed using Nexus Copy Number5.0 (BioDiscovery). The Human 1M CV HapMap control set provided by the manufacturer was used as a control. The algorithm used for the analysis was SNP-FASST Rank segmentation. Standard settings were adjusted: a significant threshold of 1.0E06,max contagious probe spacing 1000 kb, minimal number of probes per segment 15, high gain 0.6, gain 0.2, loss -0.2, big loss -1, homozygous frequency threshold 0.9, homozygous threshold 0.85, heterozygosity imbalance threshold 0.35 and minimum loss of heterozygosity (region with loss of heterozygosity) length of 5000 (kb). The plotted Log R Ratio and the B Allele Frequency by Nexus Copy Number 5.0were used to perform CNV calling. Log R Ratio is the ratio between the observed and the expected probe intensity. The expected probe intensity is an interpolation of the mean intensities of the surrounding probe clusters. The B Allele Frequency is a value between 0 and 1, which represents the proportion contributed by one SNP allele (B) to the total copy number.

Affymetrix exon-array analysis

Exon-array data from 89 PCa patient samples, 11 PCa xenografts and 6 cell PCa cell lines were collected using the GeneChip Human Exon 1.0 ST Array (Affymetrix). The staining, washing and scanning procedures were performed according to the manufacturers protocol (Affymetrix). The processing and RMA quantile normalization of the data were performed using the R-package affy (http://www.bioconductor.org/packages/release/bioc/manuals/affy/man/affy.pdf). The list of genes with DNA breakpoint originated from the SNP microarray data was analyzed to determine differential expression of exons in these genes (GSE41410). On the GeneChip Human Exon 1.0 ST Array, 5'362'207 probes (on average 40 probesets per gene) are used to analyze one million exon clusters (collections of overlapping exons). The microarray chip is based on a selection

of validated and predicted gene locations. The file containing the raw intensity values for the core probes (derived from RefSeq transcripts or full-length mRNAs) was used to assess exon expression profile.

Cell culture

The PC346C cell line was cultured in DMEM-F12 (BioWhittaker), supplemented with 2% (volume/volume) FCS (PAN Biotech), 1% insulintransferrin-selenium (GIBCO BRL), 0.01% BSA (Boehringer), 10 ng/ml epidermal growth factor (Sigma-Aldrich, Milan, Italy) and 500U penicillin–streptomycin (BioWhittaker), 100 ng/ml bronectin (Harbor BioProducts, Tebu-bio, The Netherlands), 20 mg/ml fetuine (ICN Biomedicals), 0.1 nM R1881 (Sigma-Aldrich), 50 ng/ml cholera toxin (Sigma-Aldrich), 0.1mM phosphoethanolamine (Sigma-Aldrich), 0.6 ng/ml triodothyronine (Sigma-Aldrich) and 500 ng/ml dexamethasone (Sigma-Aldrich). The PC346C cell line is an androgen-sensitive PCa cell line derived from the PC346P xenograft. It expresses the wild-type androgen receptor and secretes high levels of prostate-specific antigen. RWPE-1 cells were cultured in keratinocyte medium (GIBCO BRL), supplemented with 5 ng/ml epidermal growth factor, 1% penicillin–streptomycin (BioWhittaker) and 50 mg/l bovine pituitary extract. RWPE-1 is a normal prostate epithelium cell line that is androgen-independent and expresses both the androgen receptor and prostate-specific antigen. Both cell lines were cultured at 37 degrees with 5% carbon dioxide.

RNA and DNA isolation

Total RNA was isolated from the PC346C, RWPE-1, PC346P and G089 using the RNeasy kit (Qiagen) and according to the manufacturer's protocol. RNA was eluted in 50 ml of RNase-free water. Concentration and purity of RNA were assessed using the NanoDrop ND1000 spectrophotometer (Nanodrop products) by absorption measurements at 260 nm. RNA was stored at -80 degrees. DNA was isolated using the QIAamp DNA Blood Midi Kit (Qiagen) according to the manufacturers' instructions. Cell pellets were resuspended in 1 ml phosphate-buffered saline (BioWhittaker) and instructions were followed according to the protocol used for 1ml whole blood. DNA was eluted in 200 µl elution buffer. Concentration and purity of the DNA were assessed using the NanoDrop ND 1000 spectrophotometer (Nanodrop products) by absorption measurements at 280 nm. DNA was stored at -20 degrees.

Sequencing

Purified PCR products have been sequenced bidirectionally using standard Sanger sequencing. A sequencing PCR was carried out using the same forward and reverse primers as used during the RT–PCR except with different primer concentrations (3 ng of primer per reaction) in a 20 ml reaction volume. The PCR product was sequenced on an ABI Model 3730 automated sequencer and analyzed using DNAMAN (Lynnon Corporation).

RT–PCR

Complementary DNA (cDNA) was synthesized using 1 mg total RNA, M-MLV reverse transcriptase kit (Promega) and Oligo (dT) primer (Invitrogen) according to the manufacturer's protocol. Reverse transcription was performed at 37 degrees Celsius for 60 min and 95 degrees Celsius for 10 min. Reverse transcription PCR for multiple fusions was carried out by amplification of the cDNA or DNA samples with the HotstarTaq Kit (Qiagen). For the standard PCR reactions, 1 µl of cDNA was used in a 50 µl reaction containing 5 ml 10x PCR Buffer, 2 µl of each primer solution (100 mM), 0.25 µl HotstarTaq (5 units/ml), 1 ml of dNTPs (10mM) and 40.75 µl of nuclease-free water. An initial denaturation step of 15 min at 95 1C was used to activate the HotstarTaq, followed by 35 cycles consisting of a denaturation step at 95 1C for 30 s, an annealing step at 55 1C for 30 s and an elongation step at 72 1C for 1 min. A final elongation step at 72 1C for10 min was used. PCR products were checked using electrophoresis of20 ml of product in a 1% agarose gel. Products of the expected size were extracted from the agarose gel using the Nucleospin Extract II Kit (Macherey-Nagel), according to the manufacturer's protocol.

Real Time–PCR

Quantitative PCR was performed using SYBR Green dye and TaqMan gene expression assays on an Applied Biosystems StepOne Real Time – PCRsystem (Applied Biosystems). All reactions were performed with SYBR Green Master Mix (Applied Biosystems) and 25 ng of both the forward and reverse primer using the manufacturer's recommended thermocycling conditions. The TaqMan gene expression assays consisted of a combination of custom primers and probes targeting MPP5–FAM71D and FAM71D. Oligo probe MF — 5' TCA GGT CAA CAG AAG AGG TGA 3' - 5'FAM 3'TAMRA and primers 5' TCT CCA ACG CAC AAG ATC TT 3' (F0) and 5' CAG TTG GCT CGG TTA TGA AGG 3' (R0) were designed to target the MPP5–FAM71D fusion. Oligo probe F — 5' CTC CTG ACA TCT CCT CCT GC 3' - 5'FAM 3'TAMRA and primers 5' CTA CTG GCC CAT CTG ACA CC 3' (F0) and 5' AGG CTC ATG TTC TCC GCA T 3' (R0) were designed to target *FAM71D.* The comparative threshold cycle (Ct) method and the quantitative standard curve method were used to quantify the expression of the target genes. Equal efficiencies of the primers were confirmed using serial dilutions of PCa cDNA. Products from the PCR were resolved by electrophoresis on 1%agarose gels and sequenced as described above if necessary. Quantitative PCR expression values for the TaqMan gene expression assays were normalized by the expression of the housekeeping genes *GAPDH* and/or *PBGD*.

Functional role of *FAM71D* in the PC346C cell line

The functional role of *FAM71D* in the PC346C cell line was assessed with the knockdown of both *FAM71D* and *MPP5–FAM71D* fusion gene. Proliferation assays were performed at different time points. The small interfering RNA reagent siGENOME SMARTpool, Human *FAM71D* (ThermoFisher Scientic) interacts with a part of the *FAM71D* gene, which is also present in the fusion. Briefly, PC346C and RWPE-1 (control) cells were

plated in 96-well culture plates at a density of 2000 cells per well (100 ml) and plated in T25 tissue flasks. Outer wells of the culture plates were filled with phosphate-buffered saline in order to avoid culturing artifacts in the proliferation assays. After 2 days, cells were washed using phosphate-buffered saline and transfected using siGENOME SMARTpool, Human *FAM71D* (final concentration of 100 nM) and Dharmafect3 reagent according to the manufacturer's protocol (Thermo Scientific) in medium without P/S. As negative controls, medium only, Dharmafect with water and a transfection with non-targeting small interfering RNA were included. Twenty-four hours after transfection the medium was replaced with normal medium, without P/S and cells were allowed to grow for a maximum of 4 days. Proliferation was assessed using the CellTiter-Glo Luminescent Cell Viability Assay (Promega) according to the manufacturer's protocol. Apoptosis was assessed using the ApoLive-Glo MultiplexAssay (Promega) that measures both the number of viable cells as a marker of cytotoxicity and caspase activation as a marker of apoptosis within a single assay well. The assay was performed according to the manufacturer's protocol. As positive control of apoptosis an apoptosis inducer set (Millipore) is used. A mix containing 700x dilutions of Actinomycin D (10mM), Camptothecin (2mM), Cycloheximide (100mM), Dexamethasone (10mM) and Etoposide (10mM) was used for 24 h to induce apoptosis in both PC346C and RWPE-1 cell lines.

Acknowledgments

We would like to thank André Uitterlinden from the Department of Internal Medicine, Erasmus MC for microarray assistance, Arno van Leenders from the Department of Pathology, Erasmus MC for patient sample selection, Wytske van Weerden from the Department of Urology, Erasmus MC for the expertise in PC346C cell line model systems, Complete Genomics Inc. for assistance with the NGS data and the patients whose material was used for this study. This research was made possible by financial contributions from CTMM, project PCMM (project number 03O-203), the FP7 Marie Curie Initial Training Network PRO-NEST (grant number 238278) and the Foundation for Scientific Urological Research (SUWO).

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. Cancer J Clin 2012; 62: 10–29.
2. Shen MM, Abate-Shen C. Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev 2010; 24: 1967–2000.
3. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W et al. Recurrent fusion of TMPRSS2 and ETS Transcription factor genes in prostate cancer. Science 2005; 310: 644–648.
4. Dong J-T. Chromosomal deletions and tumor suppressor genes in prostate cancer. Cancer Metastasis Rev 2001; 20: 173–193.
5. Clegg NJ, Couto SS, Wongvipat J, Hieronymus H, Carver BS, Taylor BS et al. MYC cooperates with AKT in prostate tumorigenesis and alters sensitivity to mTOR inhibitors. PLoS One 2011; 6: e17449.
6. Cheng I, Levin AM, Tai YC, Plummer S, Chen GK, Neslund-Dudas C et al. Copy number alterations in prostate tumors and disease aggressiveness. Genes Chromosomes Cancer 2011; 51: 66–76.
7. Jin G, Sun J, Liu W, Zhang Z, Chu LW, Kim S-T et al. Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis 2011; 32: 1057–1062.
8. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol 2011; 12: 245–255.
9. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS et al. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010; 18: 11–22.
10. Vainio P, Wolf M, Edgren H, He T, Kohonen P, Mpindi J-P et al. Integrative genomic, transcriptomic, and RNAi analysis indicates a potential oncogenic role for FAM110B in castration-resistant prostate cancer. Prostate 2011, 72: 789–802.
11. Kim J, Yu J. Interrogating genomic and epigenomic data to understand prostate cancer. Biochim et Biophys Acta 2012; 1825: 186–196.
12. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010; 11: 685–696.
13. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011; 52: 413–435.
14. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet 2010; 11: 476–486.
15. Singleton AB. Exome sequencing: a transformative technology. Lancet Neurol 2011; 10: 942–946.
16. Haimovich AD. Methods, challenges, and promise of next-generation sequencing in cancer biology. Yale J Biol Med 2011; 84: 439–446.
17. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet 2011; 48: 580–589.
18. Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biol 2010; 11: R57.
19. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet 2011; 52: 413–435.

20. Asan, Geng C, Chen Y, Wu K, Cai Q, Wang Y et al. Paired-end sequencing of long-range DNA fragments for de novo assembly of large, complex mammalian genomes by direct intra-molecule ligation. PLoS One 2012; 7: e46211.
21. Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D et al. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. Genome Res 2012; 22: 125–133.
22. Yao F, Ariyaratne PN, Hillmer AM, Lee WH, Li G, Teo AS et al. Long span DNA paired-end-tag (DNA-PET) sequencing strategy for the interrogation of genomic structural mutations and fusion-point-guided reconstruction of amplicons. PLoS One 2012; 7: e46152.
23. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG et al. Human genome sequencing using unchained base reads on self-assembling DNA Nanoarrays. Science 2010; 327: 78–81.
24. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY et al. The genomic complexity of primary human prostate cancer. Nature 2011; 470: 214–220.
25. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer- associated long noncoding RNAs and aberrant alternative splicings. Cell Res 2012; 22: 806–821.
26. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011; 29: 742–749.
27. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 2012; 44: 685–689.
28. Spans L, Atak ZK, Van Nieuwerburgh F, Deforce D, Lerut E, Aerts S et al. Variations in the exome of the LNCaP prostate cancer cell line. Prostate 2012; 72: 1317–1327.
29. Lindberg J, Klevebring D, Liu W, Neiman M, Xu J, Wiklund P et al. Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. Eur Urol 2013; 63: 347–353.
30. Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci USA 2011; 108: 17087–17092.
31. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature 2012; 487: 239–243.
32. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J et al. Deep RNA sequencing analysis of read through gene fusions in human prostate adenocarcinoma and reference samples. BMC Med Genomics 2011; 4: 11.
33. Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci USA 2011; 108: 9172–9177.
34. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L et al. Comprehensive resequence analysis of a 136 Kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008; 124: 161–170.
35. Watson PA, Chen YF, Balbas MD, Wongvipat J, Socci ND, Viale A et al. Constitutively active androgen receptor splice variants expressed in castration- resistant prostate cancer require full-length androgen receptor. Proc Natl Acad Sci USA 2010; 107: 16759–16765.

36. Yeager M, Deng Z, Boland J, Matthews C, Bacior J, Lonsberry V et al. Comprehensive resequence analysis of a 97 Kb region of chromosome 10q11.2 containing the MSMB gene associated with prostate cancer. Hum Genet 2009; 126: 743–750.
37. Wu C, Wyatt AW, Lapuk AV, McPherson A, McConeghy BJ, Bell RH et al. Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. J Pathol 2012; 227: 53–61.
38. Lapuk AV, Wu C, Wyatt AW, McPherson A, McConeghy BJ, Brahmbhatt S et al. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. J Pathol 2012; 227: 286–297.
39. Hermans KG, Boormans JL, Gasi D, van Leenders GJ, Jenster G, Verhagen PC et al. Overexpression of prostate-specific TMPRSS2 (exon 0)-ERG fusion transcripts corresponds with favorable prognosis of prostate cancer. Clin Cancer Res 2009; 15: 6398–6403.
40. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al. Circos: An information aesthetic for comparative genomics. Genome Res 2009; 19: 1639–1645.
41. Sneddon TP, Church DM. Online resources for genomic structural variation. Methods Mol Biol 2012; 838: 273–289.
42. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P et al. Large-scale copy number polymorphism in the human genome. Science 2004; 305: 525–528.
43. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y et al. Detection of large-scale variation in the human genome. Nat Genet 2004; 36: 949–951.
44. Meena Kishore S, Vincent TKC, Pandjassarame K. Distributions of exons and introns in the human genome. In Silico Biol 2004; 4: 387–393.
45. Rippe V, Drieschner N, Meiboom M, Escobar HM, Bonk U, Belge G et al. Identification of a gene rearranged by 2p21 aberrations in thyroid adenomas. Oncogene 2003; 22: 6111–611.
46, Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 2004; 36: 40–45.
47. Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J et al. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Proc Natl Acad Sci USA 2012; 109: 3416–3421.
48. Funke L, Dakoji S, Bredt DS. Membrane-associated guanylate kinases regulate adhesion and plasticity at cell junctions. Annu Rev Biochem 2005; 74: 219–245.
49. Gosens I, Sessa A, den Hollander AI, Letteboer SJF, Belloni V, Arends ML et al. FERM protein EPB41L5 is a novel member of the mammalian CRB–MPP5 polarity complex. Exp Cell Res 2007; 313: 3959–3970.
50. Lee M, Vasioukhin V. Cell polarity and cancer–cell and tissue polarity as a non-canonical tumor suppressor. J Cell Sci 2008; 121: 1141–1150.
51. Straight SW, Shin K, Fogg VC, Fan S, Liu CJ, Roh M et al. Loss of PALS1 expression leads to tight junction and polarity defects. Mol Biol Cell 2004; 15: 1981–1990.
52. Wang Q, Chen X-W, Margolis B. PALS1 regulates E-cadherin trafficking in mammalian epithelial cells. Mol Biol Cell 2007; 18: 874–885.
53. Marques RB, Erkens-Schulze S, de Ridder CM, Hermans KG, Waltering K, Visakorpi T et al. Androgen receptor modifications in prostate cancer cells upon long-termandrogen ablation and antiandrogen treatment. Int J Cancer 2005; 117: 221–229

Supplementary data

Table S1: Overview of the CGI data generated for the G089 and PC346C samples

CompleteID	GS00241-DNA_D03	GS00241-DNA_E03	GS00241-DNA_F03
SampleID	RVM2-LT	G089	PC346C
Gender	MALE	MALE	MALE
GenomeSize	2.768	2.768	2.768
Coverage	62.23265896	61.1849711	67.42052023
Gross mapping yield (Gb)	172.26	169.36	186.62
SNP Transitions/transversions	2.14	2.12455	2.19367
SNP het/hom ratio	2.14	1.11279508	1.497774806
INS het/hom ratio	1.71	1.01138642	1.645551705
DEL het/hom ratio	2.57	1.249980157	2.640563175
SUB het/hom ratio	2.42	1.362113191	1.695779779
SNP total count	3881532	2980648	3409249
INS total count	214311	178001	202800
DEL total count	225761	182105	618504
SUB total count	78182	62590	81190
SNP novel rate	14.09%	6.60%	11.29%
INS novel rate	28.88%	21.49%	35.06%
DEL novel rate	33.29%	23.05%	69.04%
SUB novel rate	40.67%	34.15%	44.94%
Fully called genome fraction	95.19%	93.85%	95.96%
Partially called genome fraction	0.91%	1.50%	0.64%
No-called genome fraction	3.89%	4.66%	3.40%
Synonymous SNP loci	11306	8118	9752
Missense SNP loci	10623	8084	10542
Nonsense SNP loci	111	83	171
Nonstop SNP loci	29	22	31
Frame-shifting INS loci	126	113	188
Frame-shifting DEL loci	142	103	458
Frame-shifting SUB loci	10	20	20
Frame-preserving INS loci	116	97	112
Frame-preserving DEL loci	121	98	135
Frame-preserving SUB loci	274	238	266
Frame-shifting/preserving ratio	0.544031311	0.545034642	1.298245614
Nonsyn/syn SNP ratio	0.939589598	0.995811776	1.081009024
Insertion/deletions ratio	0.949282648	0.977463551	0.327887936
Ins+del/SNP ratio	0.113375853	0.120814668	0.240904668
Coding insertion/deletions ratio	0.735426009	0.831325301	0.416206262
Coding SNP/all SNP ratio	0.006213268	0.006192278	0.006665984
Coding (ins+del)/all (ins+del) ratio	0.000879402	0.000844196	0.000936316

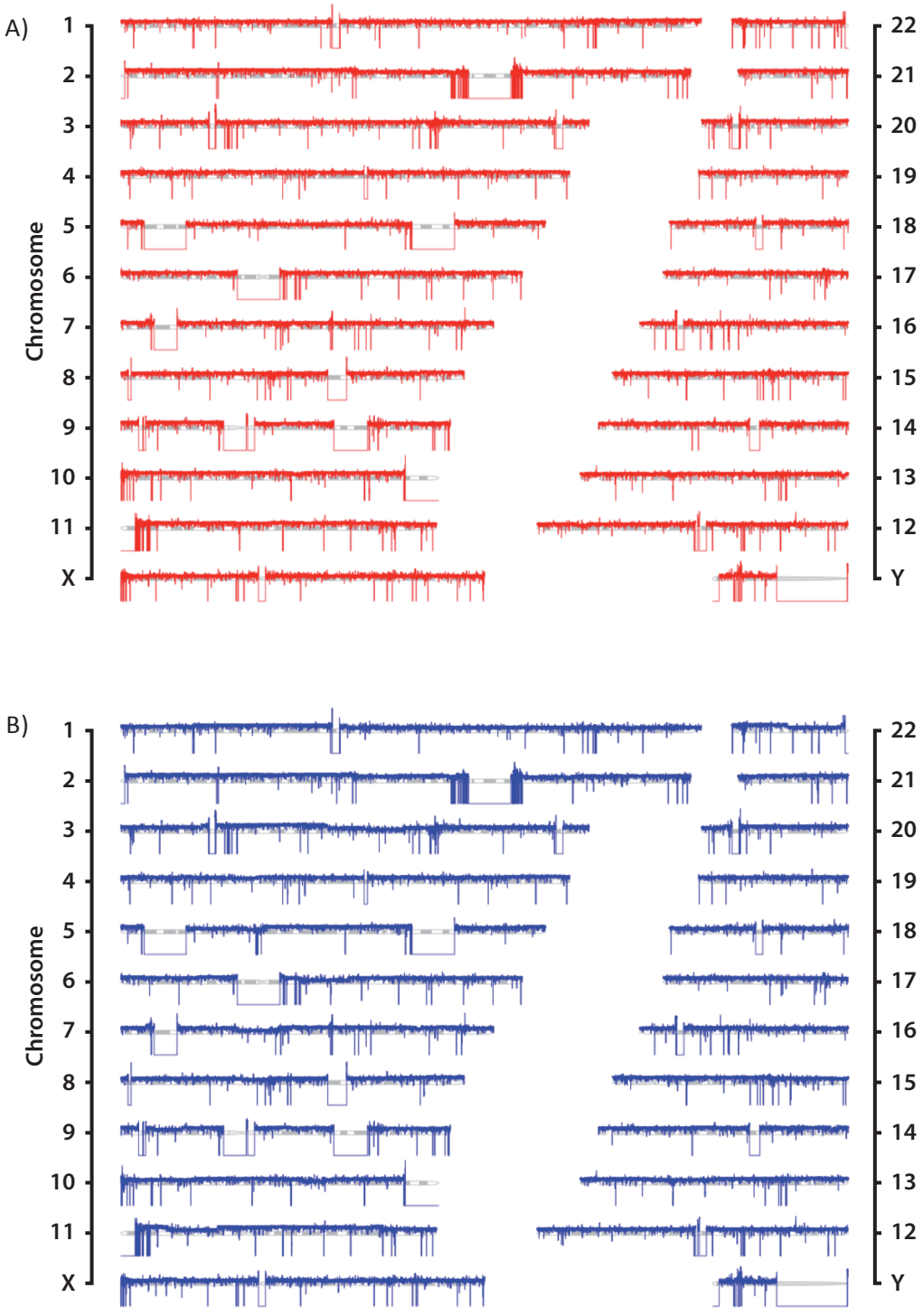


Figure S1: Coverage plots of all chromosomes in PC346C (A) and G089 (B)

Table S2: Normal genome dataset sequenced by Complete Genomics Inc.

ASW: African ancestry in Southwest USA	NA19700	LWK: Luhya in Webuye, Kenya	NA19017
	NA19701		NA19020
	NA19703		NA19025
	NA19704		NA19026
	NA19834	MKK: Maasai in Kinyawa, Kenya	NA21732
CEU: Utah residents with Northern and Western European ancestry from the CEPH collection	NA06985		NA21733
	NA06994		NA21737
	NA07357		NA21767
	NA10851	MXL: Mexican ancestry in Los Angeles, California	NA19735
	NA12004		NA19648
CHB: Han Chinese in Beijing, China	NA18526		NA19649
	NA18537		NA19669
	NA18555		NA19670
	NA18558	TSI: Toscans in Italy	NA20502
GIH: Gujarati Indian in Houston, Texas, USA	NA20845		NA20509
	NA20846		NA20510
	NA20847		NA20511
	NA20850	YRI: Yoruba in Ibadan, Nigeria	NA18501
JPT: Japanese in Tokyo, Japan	NA18940		NA18502
	NA18942		NA18504
	NA18947		NA18505
	NA18956		NA18508
			NA18517
			NA19129

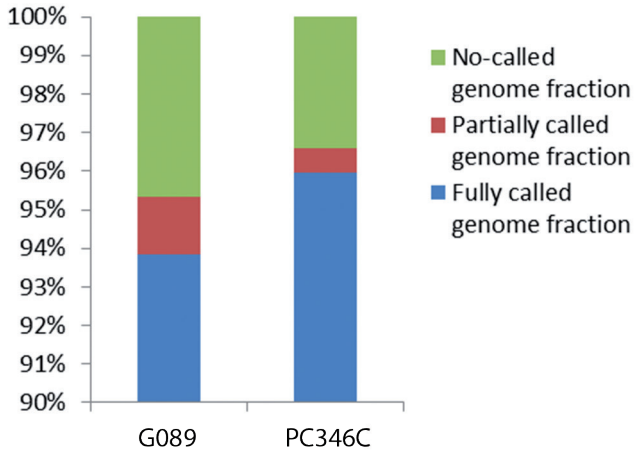


Figure S2: Structural variations assembled in G089 and PC346C

Table S3: Structural variations within genes on both junctions in the 2 prostate cancer samples corrected for SVs in control samples

type of SV	gene both sides		same gene on both sides				different gene on both sides			
			--> --> or <-- <--		--> <-- or <-- -->		--> --> or <-- <--		--> <-- or <-- -->	
	G089	PC346C	G089	PC346C	G089	PC346C	G089	PC346C	G089	PC346C
interchrom.	18	5	0	0	0	0	7	4	11	1
intrachrom.	174	101	138	82	21	10	8	9	7	0

Table S4: Structural variations within genes in the 46 normal samples

type of SV	gene both sides	same gene on both sides		different gene on both sides	
		--> --> or <-- <--	--> <-- or <-- -->	--> --> or <-- <--	--> <-- or <-- -->
interchrom.	205	0	0	114	91
intrachrom.	4329	3858	429	24	18

Table S5: Inventory of the interchromosomal SV events with genes on both sides and the intrachromosomal SV events with different genes on both sides in all the 46 normal samples (all the fusions reported have the same orientation)

LeftChr	LeftPosition	RightChr	RightPosition	Left gene	Right gene	Incidence	DiscordantMatePair Alignments
chr6	24791966	chr22	31258560	ACOT13	SYN3	6 in 46	143
chr2	114423020	chrX	44717703	ACTR3	KDM6A	1 in 46	3
chr12	4613176	chr12	7243075	AKAP3	PEX5	3 in 46	7
chr2	73642308	chr7	125890311	ALMS1	GRM8	1 in 46	101
chr4	114128776	chr6	57417655	ANK2	PRIM2	1 in 46	93
chr22	34927633	chr22	34981344	APOL4	APOL1	1 in 46	4
chr3	56740660	chr7	147170604	ARHGEF3	CNTNAP2	4 in 46	133
chr7	97334171	chr8	87780406	ASNS	CNGB3	4 in 46	25
chr12	55276363	chr15	37781916	BAZ2A	FSIP1	1 in 46	161
chr2	60598575	chr3	14519784	BCL11A	GRIP2	2 in 46	85
chr3	114462290	chr10	60051725	BOC	BICC1	2 in 46	3
chr1	109452157	chr22	28493372	C1orf194	UQCR10	1 in 46	7
chr5	175666015	chr18	53532573	C5orf25	ATP8B1	1 in 46	6
chr6	121589790	chr16	87294265	C6orf170	RNF166	2 in 46	45
chr8	96334897	chr19	18696612	C8orf37	CRTC1	2 in 46	56
chr7	44260822	chr10	60738301	CAMK2B	FAM13C	1 in 46	5
chr7	26219493	chr15	38641486	CBX3	C15orf57	6 in 46	40
chr7	80075314	chr8	25228569	CD36	DOCK5	1 in 46	4
chr1	1609907	chr1	1662846	CDK11B	SLC35E2	1 in 46	4
chr1	16249105	chr1	16258921	CLCNKB	FAM131C	3 in 46	11
chr8	88045498	chr9	117184202	CNBD1	DEC1	1 in 46	4
chr1	86171219	chr9	99715354	COL24A1	C9orf156	5 in 46	67
chr1	33896536	chr20	40323470	CSMD2	PTPRT	1 in 46	67
chr12	56506829	chr15	36538458	CTDSP2	FAM98B	5 in 46	8
chr19	46041457	chr19	46073355	CYP2A6	CYP2A7	1 in 46	3
chr1	85756077	chr1	210549248	DDAH1	PPP2R5A	1 in 46	3
chr11	108090957	chr13	20640540	DDX10	SKA3	1 in 46	14
chr11	84872658	chr17	30502225	DLG2	UNC45B	1 in 46	3070
chr11	84872657	chrX	47634658	DLG2	ZNF81	2 in 46	4
chr10	128606023	chr13	44660068	DOCK1	GTF2F2	1 in 46	4
chr2	116093152	chr9	130498485	DPP10	SET	3 in 46	15
chr2	115411737	chr17	30502225	DPP10	UNC45B	2 in 46	129
chr2	62811531	chr14	66581210	EHBP1	GPHN	2 in 46	10
chr14	66903963	chr14	66932740	EIF2S1	PLEK2	3 in 46	35
chr13	42967826	chr15	39638687	ENOX1	TYRO3	1 in 46	14
chr12	1463516	chr12	20577913	ERC1	PDE3A	2 in 46	7
chr6	133635641	chr16	46096510	EYA4	PHKB	2 in 46	5
chr2	70377092	chr2	179004389	FAM136A	PRKRA	1 in 46	3
chr2	75587805	chr7	103207935	FAM176A	RELN	1 in 46	4
chr4	46662762	chr22	27395456	GABRA4	TTC28	1 in 46	31
chr1	89371762	chr1	89424837	GBP7	GBP4	1 in 46	21
chr7	2748187	chr7	26189704	GNA12	NFE2L3	1 in 46	5
chr14	24146348	chr14	24171060	GZMH	GZMB	1 in 46	8

Table S5: Continued

LeftChr	LeftPosition	RightChr	RightPosition	Left gene	Right gene	Incidence	DiscordantMatePair Alignments
chr1	91625370	chr7	48642900	HFM1	ABCA13	2 in 46	37
chr1	91625736	chr10	50581095	HFM1	C10orf53	1 in 46	3
chr1	91625370	chr4	108144352	HFM1	DKK2	2 in 46	4
chr1	91625736	chr11	77275449	HFM1	INTS4	1 in 46	5
chr1	91625370	chr12	84086888	HFM1	LRRIQ1	1 in 46	11
chr1	91625370	chr22	20540553	HFM1	MAPK1	2 in 46	10
chr1	91625370	chr2	159149043	HFM1	PKP4	1 in 46	50
chr1	91625370	chr4	77026210	HFM1	PPEF2	2 in 46	245
chr1	91625736	chr1	235833345	HFM1	RYR2	2 in 46	12
chr1	91625370	chr17	72669985	HFM1	SEC14L1	3 in 46	88
chr1	91625736	chr8	70764802	HFM1	SLCO5A1	1 in 46	18
chr1	91625370	chr7	104753533	HFM1	SRPK2	1 in 46	11
chr1	91625736	chr6	144733889	HFM1	UTRN	1 in 46	4
chr6	32600654	chr6	32660804	HLA-DRB5	HLA-DRB1	2 in 46	3
chr3	123981035	chr10	17752794	HSPBAP1	STAM	1 in 46	7
chr2	149355896	chr5	79983676	KIF5C	DHFR	1 in 46	6
chr17	36760562	chr17	36779545	KRT33A	KRT33B	2 in 46	9
chr12	51130451	chr12	51171567	KRT6B	KRT6A	1 in 46	4
chr7	55414342	chr12	4814579	LANCL2	KCNA6	1 in 46	6
chr1	26610053	chr2	179030761	LIN28A	DFNB59	1 in 46	3
chr6	119600390	chr19	40838511	MAN1A1	COX6B1	1 in 46	16
chr6	119600398	chr10	35350401	MAN1A1	CUL2	10 in 46	27
chr6	119600259	chr17	28173963	MAN1A1	MYO1D	5 in 46	40
chr6	119600395	chr8	69051583	MAN1A1	PREX2	5 in 46	23
chr6	74212891	chr14	70550021	MB21D1	PCNX	1 in 46	3
chr3	184515406	chr8	66742065	MCF2L2	MTFR1	1 in 46	4
chr2	227902744	chr5	149292455	MFF	PDE6A	1 in 46	8
chr6	31485269	chr6	31579996	MICA	MICB	1 in 46	4
chr6	43763511	chr9	33120548	MRPS18A	B4GALT1	1 in 46	16
chr2	42687780	chr10	72118662	MTA3	ADAMTS14	1 in 46	4
chr11	76581513	chr19	14567841	MYO7A	CLEC17A	7 in 46	14
chr3	176583069	chr7	96161255	NAALADL2	SHFM1	1 in 46	3
chr1	143391395	chr1	143665970	NBPF9	PDE4DIP	1 in 46	3
chr6	126274506	chr8	48655786	NCOA7	KIAA0146	1 in 46	28
chr6	126280035	chr20	33881165	NCOA7	PHF20	1 in 46	24
chr17	26575403	chr19	19493427	NF1	NDUFA13	1 in 46	3
chr3	175093152	chr20	40323474	NLGN1	PTPRT	1 in 46	109
chrX	6148397	chrY	15152001	NLGN4X	NLGN4Y	4 in 46	7
chr1	120345451	chr6	133635756	NOTCH2	EYA4	8 in 46	27
chr1	120345412	chr16	46096541	NOTCH2	PHKB	7 in 46	33
chr1	143988802	chr6	133635754	NOTCH2NL	EYA4	6 in 46	29
chr1	143988894	chr16	46096578	NOTCH2NL	PHKB	8 in 46	12
chr1	198334303	chr2	23551752	NR5A2	KLHL29	2 in 46	51
chr11	47763216	chr15	73654668	NUP160	PTPN9	1 in 46	3
chr12	20595770	chr14	89411235	PDE3A	EFCAB11	1 in 46	3

Table S5: Continued

LeftChr	LeftPosition	RightChr	RightPosition	Left gene	Right gene	Incidence	DiscordantMatePair Alignments
chr1	66397553	chr11	8136900	PDE4B	RIC3	1 in 46	4
chr5	32141959	chr5	32203532	PDZD2	GOLPH3	1 in 46	21
chr12	7247092	chr20	40997762	PEX5	PTPRT	1 in 46	3
chr6	13299425	chr17	62067736	PHACTR1	PRKCA	1 in 46	121
chr7	621590	chr10	128606010	PRKAR1B	DOCK1	1 in 46	5
chr1	212723427	chr5	115205559	PTPN14	AP3S1	2 in 46	6
chr15	73654668	chr17	62977033	PTPN9	PITPNC1	1 in 46	4
chr14	72621755	chr14	72678268	RBM25	PSEN1	5 in 46	13
chr12	55275617	chr15	37781922	RBMS2	FSIP1	1 in 46	197
chr3	124135562	chr12	48178954	SEMA5B	SPATS2	3 in 46	4
chr10	82333232	chr17	8802559	SH2D4B	PIK3R5	2 in 46	80
chr1	115124725	chr1	216013856	SIKE1	SPATA17	1 in 46	26
chr20	1540516	chr20	1843589	SIRPB1	SIRPA	1 in 46	6
chr14	36701360	chr14	36840980	SLC25A21	MIPOL1	1 in 46	6
chr1	9044032	chr14	92782930	SLC2A5	BTBD7	5 in 46	112
chr6	88274265	chr20	51455050	SLC35A1	TSHZ2	2 in 46	38
chr2	213879355	chr7	28346689	SPAG16	CREB5	2 in 46	6
chr2	213879333	chr6	31585828	SPAG16	MICB	1 in 46	3
chr3	137710453	chr16	14488371	STAG1	PARN	4 in 46	13
chr7	23727175	chr8	52895822	STK31	PCMTD1	1 in 46	6
chr8	133989904	chr14	50567849	TG	TRIM9	1 in 46	4
chr1	15390838	chr2	227930141	TMEM51	MFF	1 in 46	8
chr2	27380758	chr3	196408092	TRIM54	C3orf21	1 in 46	5
chr15	39640800	chr15	74340792	TYRO3	ETFA	1 in 46	3
chr3	48622068	chr8	52818266	UQCRC1	PXDNL	1 in 46	11
chr1	21898605	chrX	1575602	USP48	P2RY8	1 in 46	5
chr12	94204809	chr15	38641471	VEZT	C15orf57	1 in 46	5
chr5	37745606	chr7	8629826	WDR70	NXPH1	4 in 46	57
chr1	68401586	chr10	126830121	WLS	CTBP2	1 in 46	3
chr7	63927371	chr19	20921389	ZNF138	ZNF85	1 in 46	5
chr12	132289429	chr20	44564610	ZNF268	ZNF334	3 in 46	6
chr10	38350357	chr10	42443852	ZNF33A	ZNF33B	3 in 46	12
chr19	58628530	chr19	58665341	ZNF761	ZNF813	1 in 46	6

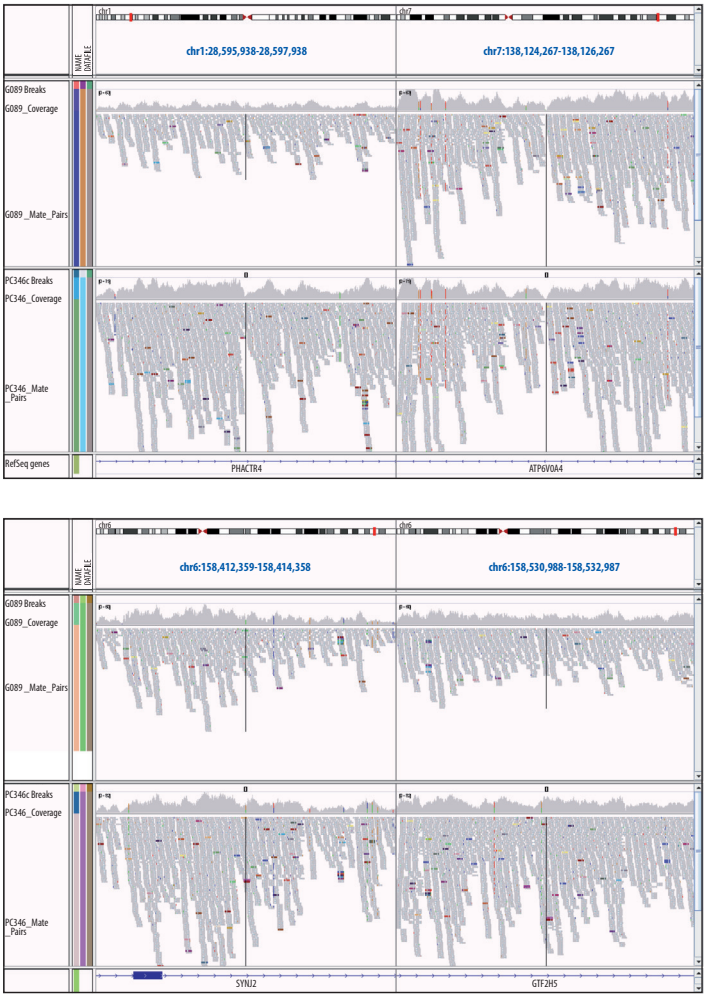


Figure S3: Mapped reads of the validated gene fusions. The upper and lower part of each of the outputs corresponds to the PC346C and G089 sample respectively. The grey reads represent the mapped reads according to the reference genome. Colored reads are discordant mate pairs mapping to different genomic locations.

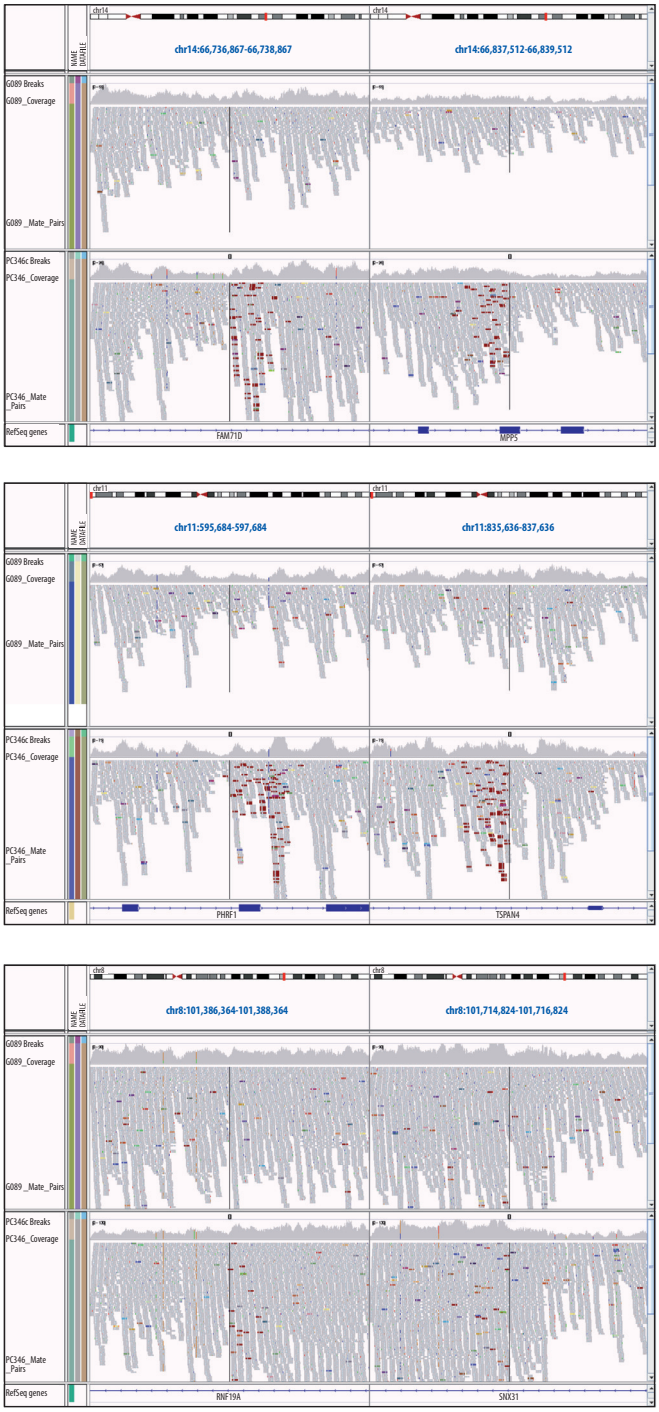


Figure S3: Continued

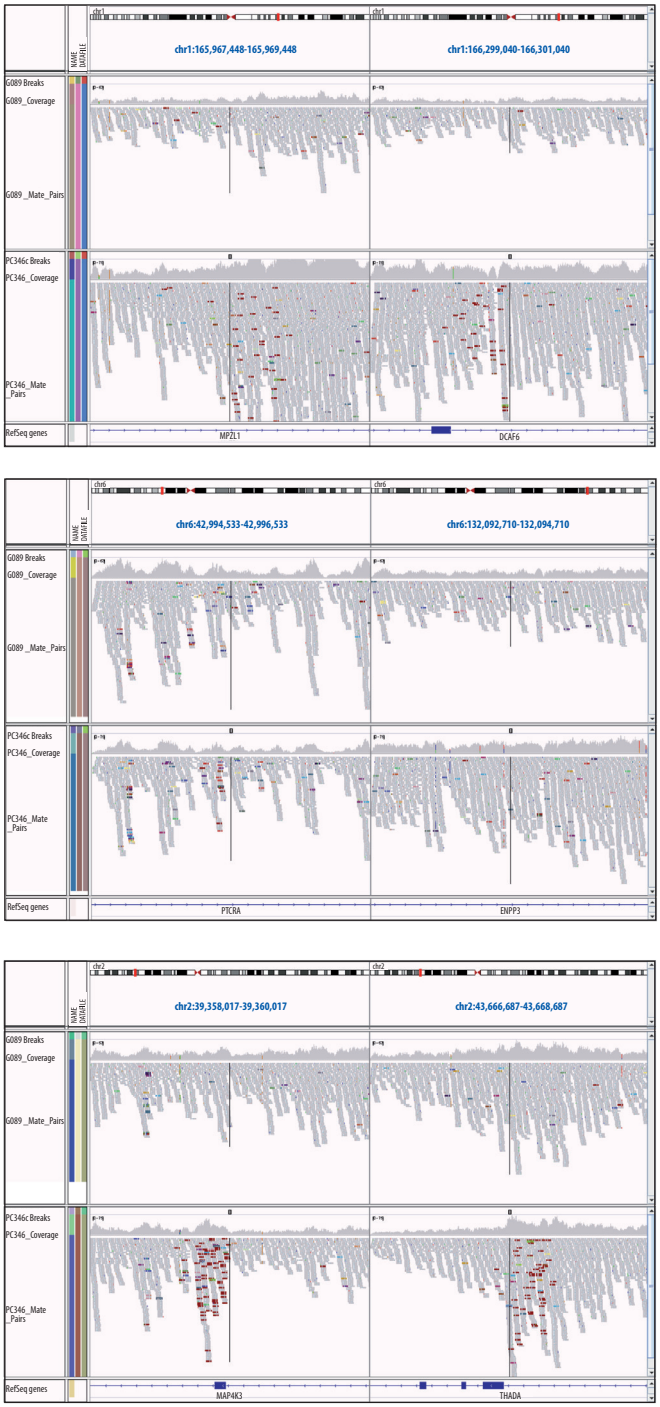


Figure S3: Continued

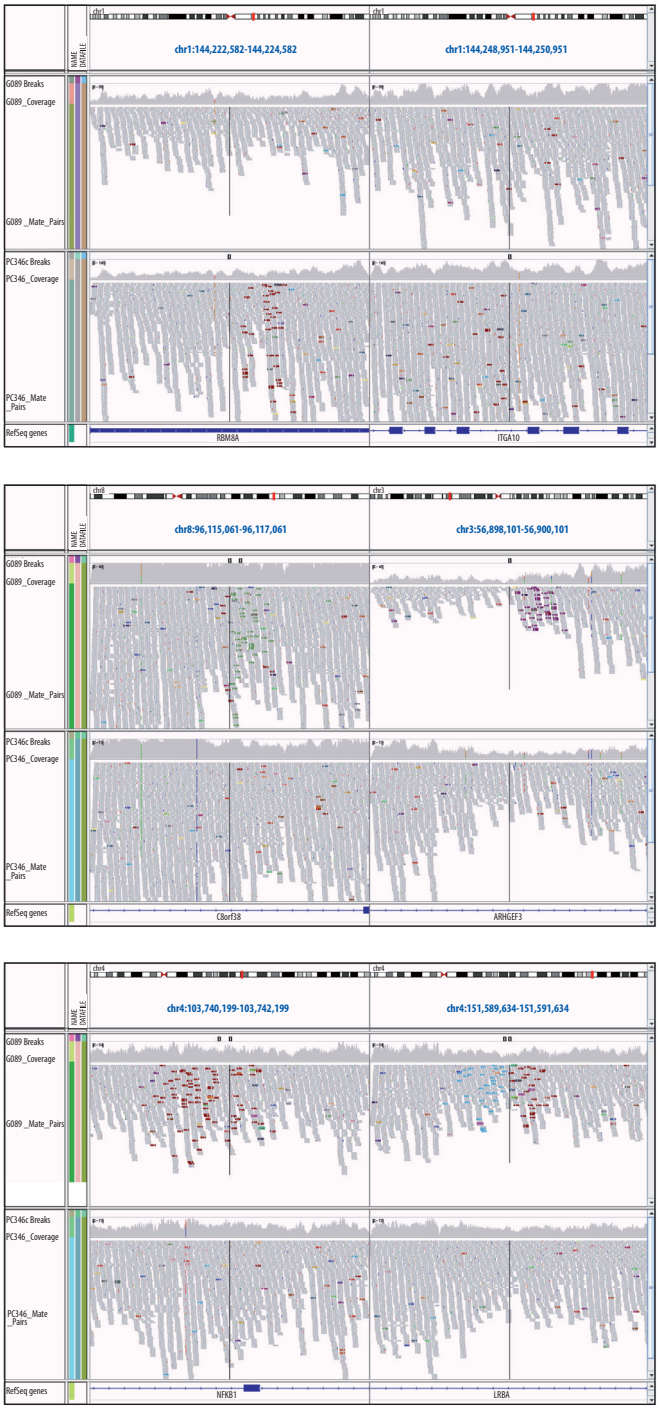


Figure S3: Continued

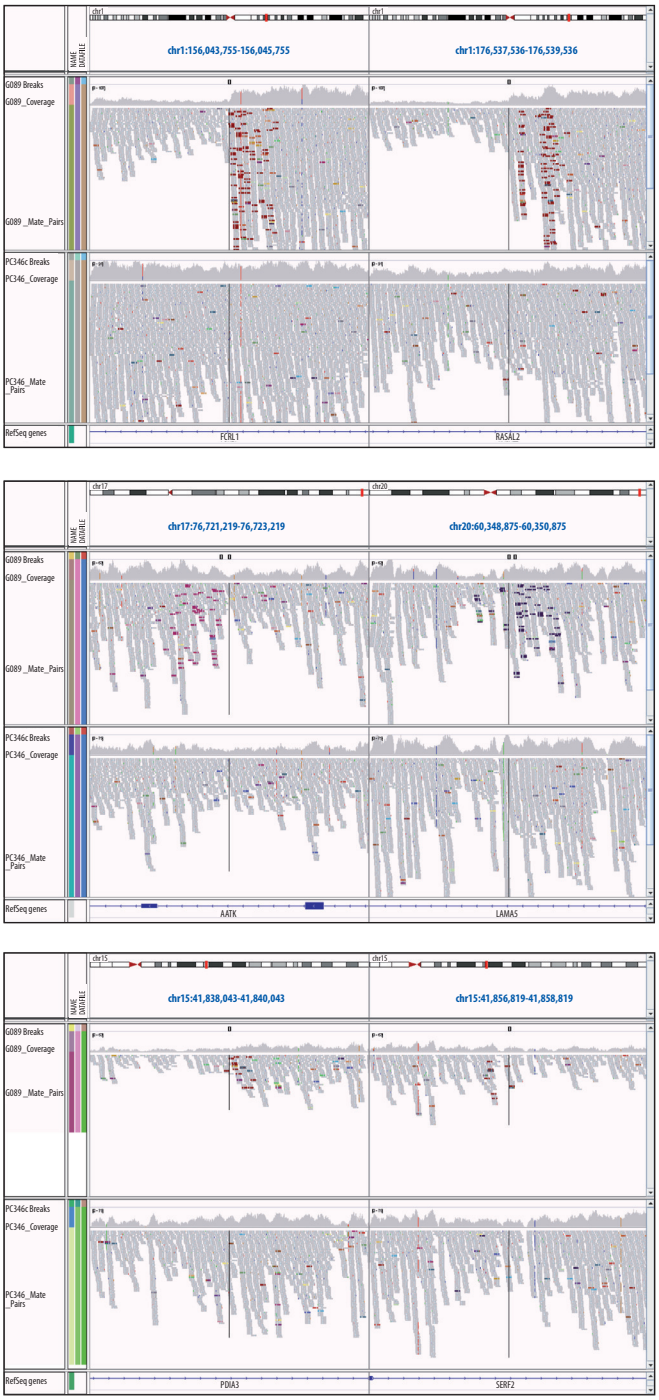


Figure S3: Continued

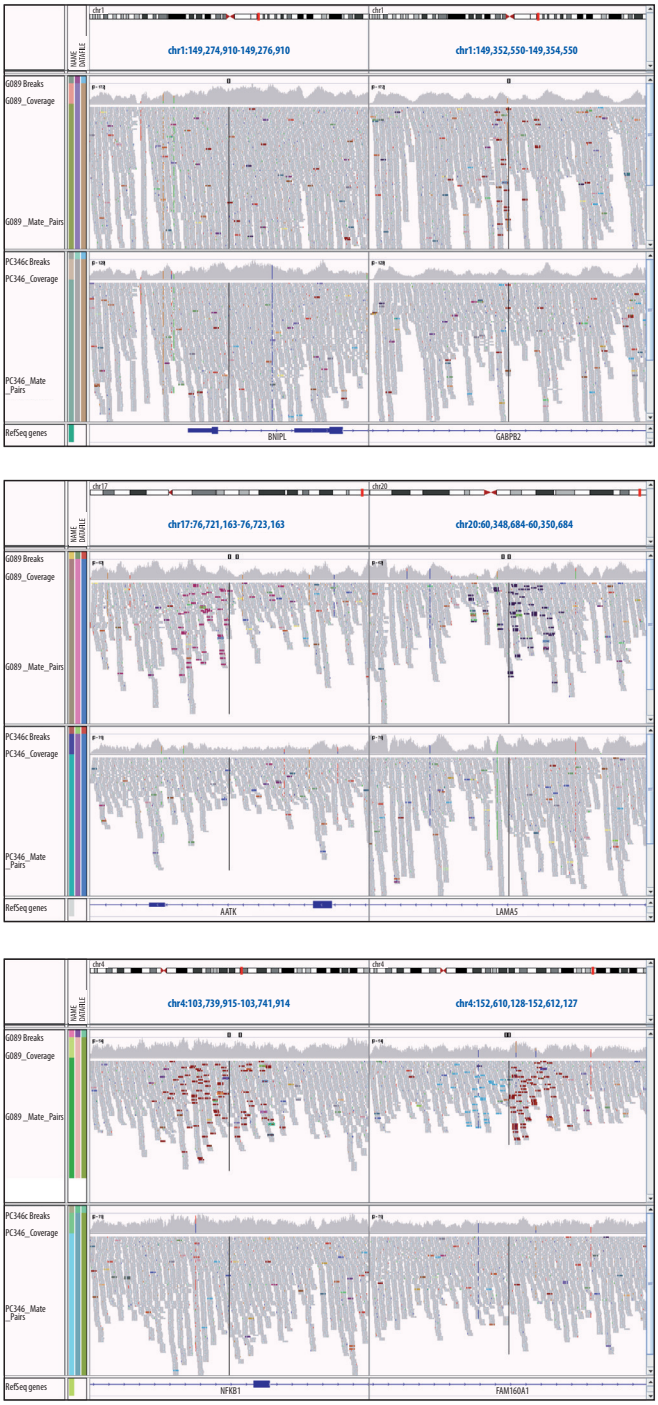


Figure S3: Continued

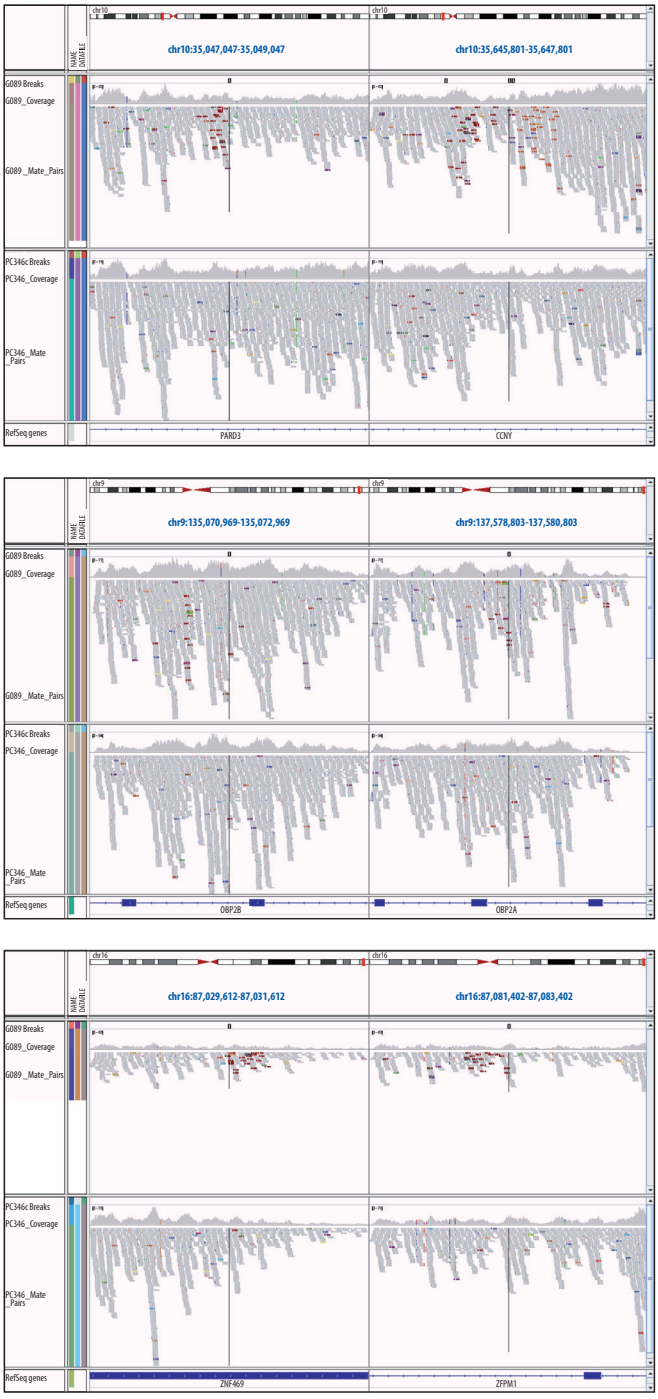


Figure S3: Continued

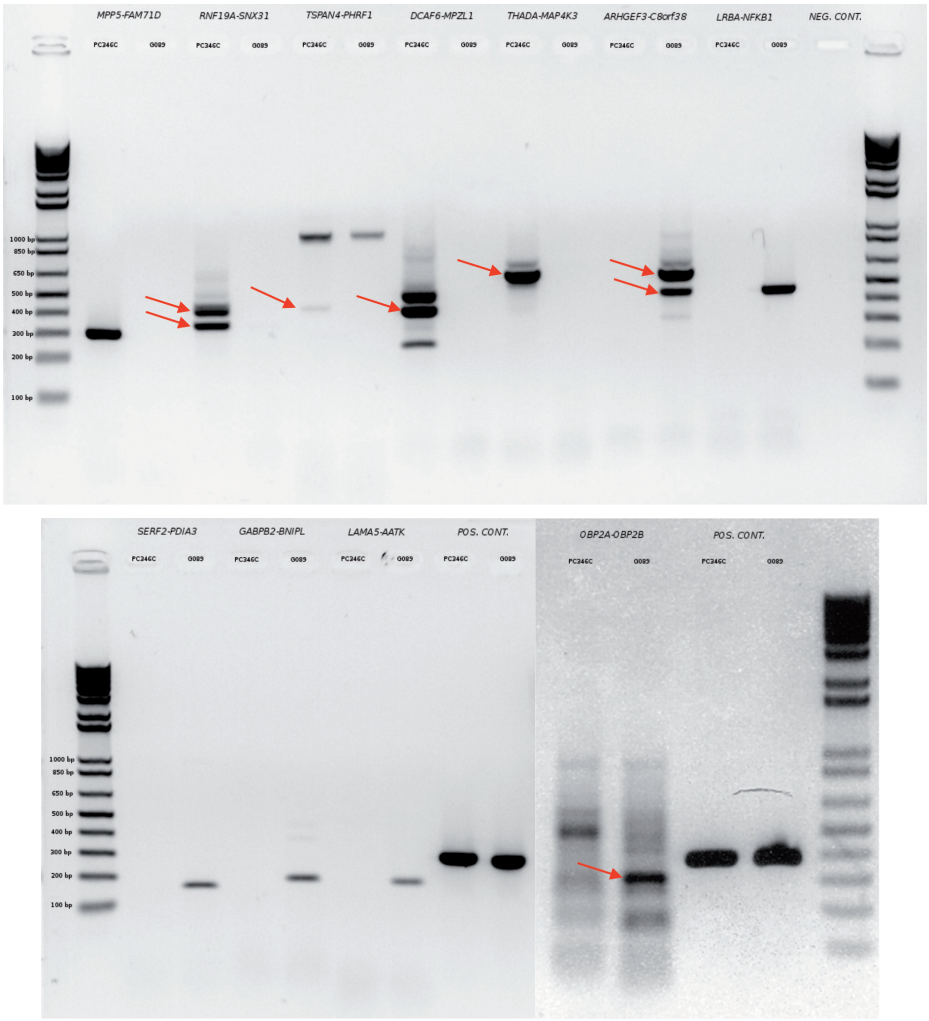


Figure S4: RT-PCR of the validated RNA fusions of PC346C and G089 (electrophoresis on a 1% agarose gel).

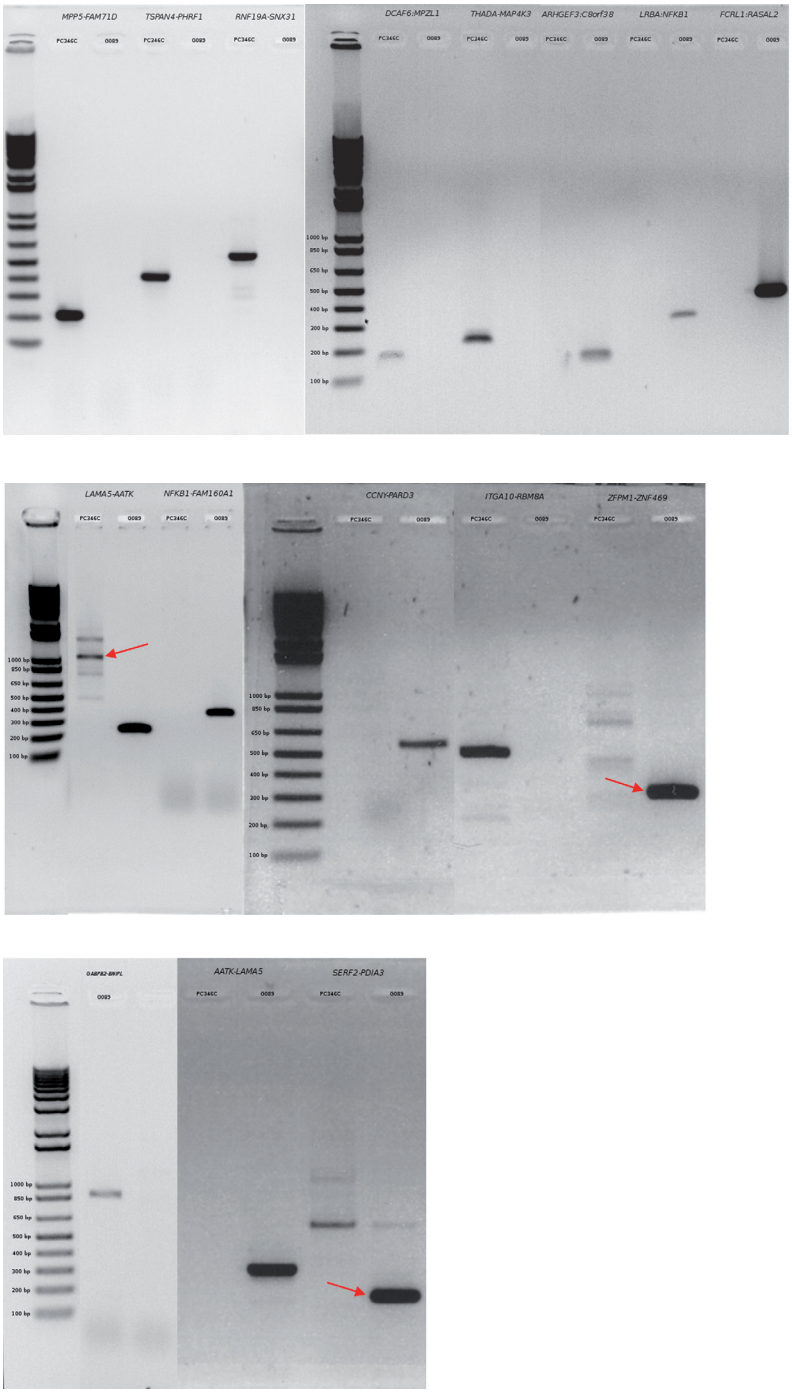


Figure S5: RT-PCR of the validated DNA fusions of PC346C and G089 (electrophoresis on a 1% agarose gel)

Table S6: Primers for the validation of the fusions on RNA level (PC346C sample)

1	TG Fw	AAGGAAGCCACCGTCATGTGCG	389
	TRIM9 Rv	TTCACTCCTAAACGCTCCCTGGTC	
2	ATP6V0A4	AGTGGAGGGAGGCAAAGG	535
	PHACTR4	CAGCAATGACAGGGTTGCTA	
3	MPP5 Fw	TGGTAGATTCTCAGAGCCAGGAGGA	369
	FAM71D Rv	ACGGTCACCCAGTTGGCTCG	
4	TSPAN4 Fw	AGAGCCAGGTTCAGCTGCGA	400
	PHRF1 Rv	GGCAAGTCCAGCCTCACAGGC	
5	RNF19A Fw	TTCGGGAAAGCGGCCGTCG	394
	SNX31 Rv	AGCCCAAGAGCTCTCGACACAGTC	
6	GTF2H5 Fw	CATCAGTAAATCTGTACTGCCCGAG	365
	SYNJ2 Rv	TCTAGAAGGTTGAGTTCTCCAGCTG	
7	DCAF6 Fw	AGACGCTCTGCTGTTGCCCGTA	389
	MPZL1 Rv	AAGGTCTCCAGCCCAGCTGATTCTG	
8	PTCRA Fw	CCGGTACATGGCTGCTACTT	175
	ENPP3 Rv	AGCCGTGGGTGATATTCTTG	
9	HLA-DRB1Fw	TGGTCCTGTCCTGTTCTCCAGCATG	329
	HLA-DRB5 Rv	CGGCGCGCCTGTCTTCCAGG	
10	CTBP2 Fw	CATTTTAAGGTGGCTCGGGAGCGG	282
	WLS Rv	TTCAGTCCACTCAGCAAACGCGT	
11	THADA Fw	TGCTGCGCTGACCATTTGCCAT	437
	MAP4K3 Rv	TTCTGGCACCATTTGGTTTCAGGG	
12	CBX3 Fw	GGTGTTGGCAGTTTAGGACCTGCT	197
	C15orf57 Rv	TTGGCTTGGGCCAGAGTTCGAA	

Table S7) Primers for the validation of the fusions on RNA level (G089 sample)

1	ARHGEF3 Fw	AACGGCAGCGGCAAAACAACTT	570
	C8orf38 Rv	GCAAGTGTGCTTGACTGGCAATGTC	
2	LRBA Fw	TTCAACTTCCCAGACCCTGCAACAG	452
	NFKB1 Rv	CCAAGTGCAAGGGCGTCTGGTAC	
3	FCRL1 Fw	ATCCCTGACCTGGTCCTCAT	335
	RASAL2 Rv	TGACTGGCTTTTGGTCCTCT	
4	AATK Fw	CGCCTTCAGCTCGCACTTCGA	377
	LAMA5 Rv	CAGGCTGGCATAGGCAGCGG	
5	SERF2 Fw	AGAGGATGGCTGACACGAAG	152
	PDIA3 Rv	GTCCTGAGAGGCACTGAAGC	
6	GABPB2 Fw	CCAGCAAGTAATGGGGAGTG	189
	BNIPL Rv	ATGTCTCAGGGCTGCTCCTA	
7	LAMA5 Fw	GAGGACCTGACGGGTCGATGC	183
	AATK Rv	TCGTCCCCCTCCGCATTCTCAAAC	
8	DOCK1Fw	GCGGCCTAGACGCGGAGTTTCC	161
	GTF2F2 Rv	TTTCACTGGCAGCTGGTCGGCAT	
9	NFKB1 Fw	ACTACGCGGTGACAGGAGAC	115
	FAM160A1 Rv	TGCTGGAAACCCCTAGAAGA	
10	CBX3 Fw	ACTGCCATCACAGCAGGTTTCCT	399
	C15orf57 Rv	TCGTCCTCGGCCACTGGCTT	
11	CCNY Fw	GGAAATCGTGCAGTGAACCT	325
	PARD3 Rv	TCCTGCTCATCAAACACTGC	
12	OBP2A Fw	CACCCTGGAGGAGGAGGATGAGG	232
	OBP2B Rv	AGAGTCCCTTGCGCTGCACC	
13	ZFPM1 Fw	GAGACATGGAGGCCAGAGAG	188
	ZNF469 Rv	GCACAGGTCTTGCATCTCAA	
14	C15orf57 Fw	CGCTGACTCTACAGCCACAA	416
	CBX3 Rv	AATCCCTTCCACTTCAGGAAA	

Table S8: Primers for the validation of the fusions on DNA level (PC346C sample)

1	TG Fw G	TCTCCCCAAAGAGCAAAGAGGTGAT	591
	TRIM9 Rv G	TGTAGGTACACTGTGGTGGCCATTT	
2	PHACTR4 Fw G	AACACAGAAGCATGTACTGCAGGCT	481
	ATP6V0A4 Rv G	CCTGACAGAGCCAGGAGCAAGAAGA	
3	FAM71D Fw G	AGCTGATGCCCTGGGGCTCT	213
	MPP5 Rv G	AACAGGCACTTTTACTGGCCCACG	
4	PHRF1 Fw G	GAGCCGTCGCGGTGGATGATGA	409
	TSPAN4 Rv G	GGCTCTATGGTTCCCTCCTGCCCTT	
5	RNF19A Fw G	GGGAGACCCCTGAGGATAAA	395
	SNX31 Rv G	CATGGCTTCCTGGTAAGGAC	
6	SYNJ2 Fw G	TGCCTTGAGTGCTCGTGCAGC	446
	GTF2H5 Rv G	TTGCAAAAATCTGTGGTGCAGCTTG	
7	MPZL1 Fw G	GTGCAGTGGCATAAACATGG	147
	DCAF6 Rv G	TTTTGGAGAGAAGGATTTGGA	
8	PTCRA Fw G	AAATAACATTCTGGGCCAGGT	128
	ENPP3 Rv G	GAAACAGGGTTTTGCCATGT	
9	HLA-DRB5 Fw G	ACTGTCACTGTGGCTTGCAT	254
	HLA-DRB1 Rv G	TCTATCTGGGCATGTGTTCA	
10	WLS Fw G	TGTCCAATTGCTGTCTCTTGA	200
	CTBP2 Rv G	GGTAGGAGCCTCGAGACCA	
11	MAP4K3 Fw G	TCTTCGGCACCAAATATCAA	174
	THADA Rv G	GCCAGGAAATTCGAGTGATG	
12	CBX3 Fw G	GTTCAGCAAAAGCCAGGAAG	171
	C15orf57 Rv G	ATACTATGGCCGCAGGACAG	
13	RBM8A Fw G	TCAAAACAAGGCTGTGGACCAAACT	503
	ITGA10 Rv G	CAAGCCCTTGCGGCTGCTCA	

Table S9: Primers for the validation of the fusions on DNA level (G089 sample)

1	ARHGEF3 Fw G	AGGTAAATGCCTTGTGGGTCT	154
	C8orf38 Rv G	ACGCAAAGAAAGGCTATGGA	
2	NFKB1 Fw G	GAATGAAGGTGGATGATTGCT	233
	LRBA Rv G	CCAGGAGGGCATTATTTTGA	
3	FCRL1 Fw G	CATAGCTCCCCATCACACCT	427
	RASAL2 Rv G	CCACACTCGAGACAGGGTTT	
4	AATK Fw G	TCAATCTGCAAAGGGGAGTC	369
	LAMA5 Rv G	GTGGGCGCCTATAATCAGTC	
5	PDIA3 Fw G	TGCTGCTGTGGATATAGGTG	292
	SERF2 Rv G	CCATGAATCAGGTTGCAGTG	
6	BNIPL Fw G	CAAGTGGCTGGAACCCAGCTAG	583
	GABPB2 Rv G	GCACACCAACCTGGGTAACAGAG	
7	AATK Fw G	GGTTTGCAGGGGGAGGAG	157
	LAMA5 Rv G	TGGGGTTTTATTTTCTTTCAACC	
8	DOCK1 Fw G	CTGCCTAGAGAGTGCGGTGGTG	600
	GTF2F2 Rv G	CATGAACAGGTCTCCCCAAAATGCAC	
9	NFKB1 Fw G	TCTGGGCTGGAAGCCTCAGTGTG	369
	FAM160A1 Rv G	CTTGGCAGCAGAAAGAAAAACCACC	
10	CBX3 Fw G	GTTCAGCAAAAGCCAGGAAG	171
	C15orf57 Rv G	ATACTATGGCCGCAGGACAG	
11	PARD3 Fw G	GGTGGCAGGTGCCTTAATC	559
	CCNY Rv G	CATGCAGGGCTAAGTGCTATC	
12	OBP2B Fw G	GACCCTCCTACACCCGAGAC	410
	OBP2A Rv G	GGGCACTGAAGACCCATCT	
13	ZNF469 Fw G	TCGCCGTGCCCTCGAAGTCTAA	252
	ZFPM1 Rv G	ATGGGCGCGATAACACAGCAGCG	
14	CBX3 Fw G	CGTGTGGCATCTGAAGCA	223
	C15orf57 Rv G	CGTGACTACCCGAACAGCTC	

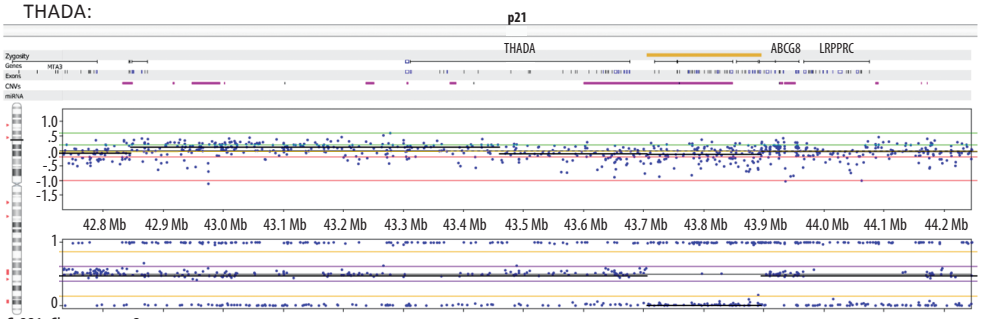
Table S10: RNA alignment of the validated gene fusions for PC346C and G089. The sequencing data was used to predict the gene fusions at RNA level. The RNA sequences and the corresponding protein translations of the first 30 bp before and after the junctions obtained by sequencing of the PCR products are depicted below. * represents the fusions in frame.

* MPP5 (ENST00000261681) Exon 5	FAM71D (ENST00000311864) Exon 4
ACTGCTTTGCTGAATACTCCACATATTCAG -T--A--L--L--N--T--P--H--I--Q-	GTCAACAGAAGAGGTGAATCCATTTACCTT -V--N--R--R--G--E--S--I--Y--L-
TSPAN4 (ENST00000397406) Exon 1	PHRF1 (ENST00000264555) Exon 13
TCAGCTGCGAGGCTGCTGGTGGCTGTGGAG	GAGGCGCCTCCCTGCCGCGGTGCCAGAGCC --R--R--L--P--A--A--V--P--E--P
RNF19A (ENST00000341084) Exon 1	SNX31 (ENST00000311812) Exon 4
CTGCGCTGCCGCCGCTCTGAGGAGTTAAAG	TAACCATGGACCCAAACGTGTTGAGAAGTG V--T--M--D--P--N--V--L--R--S--
RNF19A (ENST00000341084) Exon 1	SNX31 (ENST00000311812) Exon 5
CTGCGCTGCCGCCGCTCTGAGGAGTTAAAG	AATACATTTGACATCGCCACCAAGAAAGCT -N--T--F--D--I--A--T--K--K--A-
DCAF6 (ENST00000312263) Exon 15	MPZL1 (ENST00000359523) Exon 2
TATAAAGGCCATCGCAACTCCAGGACAATG -Y--K--G--H--R--N--S--R--T--M-	TGACAGCTGGAGTATCAGCCTTGGAAGTAT L--T--A--G--V--S--A--L--E--V--
THADA (ENST00000405975) Exon 4	MAP4K3 (ENST00000263881) Exon 24
AAAGAATCCCTTGAAGAAAGTATTGGCAAG --K--N--P--L--K--K--V--L--A--S	ATCAGTACTTGATATTTGGTGCCGAAGAAG D--Q--Y--L--I--F--G--A--E--E--
* ARHGEF3 (ENST00000338458) Exon2	C8orf38 (ENST00000396124) Exon 3
CCCTCTCCTAAAGCCTGGAATTTCAGAGGG -P--S--P--K--A--W--N--F--R--G-	GTTAAAGACTCAGTCTCTGAGAAAACAATT -V--K--D--S--V--S--E—K--T--I-
* ARHGEF3 (ENST00000338458) Exon 2	C8orf38 (ENST00000396124) Exon 4
CCCTCTCCTAAAGCCTGGAATTTCAGAGGG -P--S--P--K--A--W--N--F--R--G-	GCTGTTAAAAGACATAATCTGACTAAAAGA -A--V--K--R--H--N--L--T--K--R-
LRBA (ENST00000357115) Exon 45	NFKB1 (ENST00000394820) Exon 16
TTGCCCACCAACTTCAGAGATTTGTCCAAG -L--P--T--N--F--R--D--L--S--K-	TGTCTTACACTTAGCAATCATCCACCTTCA --V--L--H--L--A--I--I--H--L--H
SERF2 (ENST00000381359) Exon 1	PDIA3 (ENST00000300289) Exon 4
ACCGCTCTGAAACCGATTTCTAACAGAGAG	ATGGAATTGTCAGCCACTTGAAGAAGCAGG D--G--I--V--S--H--L--K--K--Q--
GABPB2 (ENST00000368918) Exon 7	BNIPL (ENST00000368931) Exon 2
TTATTGTAACTGTGCAAGATGGACAGCAAG F--I--V--T--V--Q--D--G--Q--Q--	GGTCAGGGAGATTGCAGAAGCACCAGAACT --V--R--E--I--A--E--A--P--E--L
LAMA5 (ENST00000252999) Exon 11	AATK (ENST00000326724) Exon 3
AGCAGGTGCTGCCAGCCGGCCAGATTGTGA E--Q--V--L--P--A--G--Q--I--V--	GAGTTTGAGAATGCGGAGGGGGACGAGTAC -E--F--E--N--A--E--G--D--E--Y-

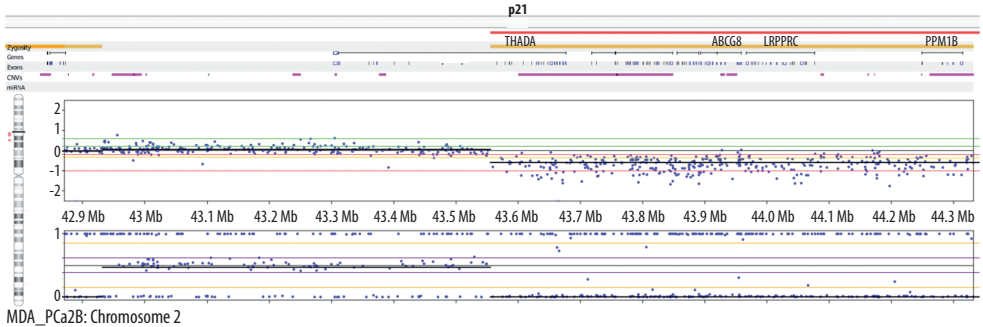
Table S11: DNA alignment of the validated gene fusions for PC346C and G089. The DNA sequences corresponding to the first 30 bp before and after the junctions obtained by sequencing of the PCR products are depicted below.

MPP5 (ENST00000261681) Exon 6	FAM71D (ENST00000311864) Intron 3-4
TGCAGCTAGAGCCCATTACAGATGAGAGAG	TAGGATGGCGCTACTGCACTCCAGCCTGGA
TSPAN4 (ENST00000397397) Intron 1-2	PHRF1(ENST00000264555) Intron 12-13
AGGAGGAGGACAGACTCCCCAGCCTGGCAC	CCTGGGTGCGGGCCCTCAGGCCGTGGGAG
RNF19A (ENST00000341084) Intron 1-2	SNX31 (ENST00000311812) Intron 3-4
GATAAATATTATATAACGTATTAATATAAA	TAATTTTTGTATTTTTTTAGTAGAGATGGG
ATAAAAAAAGAGAACATTAAGCAGGGCATTAATATAAAAATTAATA	
DCAF6 (ENST00000312263) Intron 15-16	MPZL1 (ENST00000359523) Intron 1-2
TGATTGCAAAAAATTGTC	AAATATTGCCTGGGCAACATGGCAAAACTG
THADA (ENST00000405975) Intron 4-5	MAP4K3 (ENST00000263881) Intron 23-24
TGGGGCTATTAGATTGTCAGATTTTTAATA	ATCAGTCAGTGAATTACTTTTATTTCAGAT
ATTTAAAAATATTTTATTT	
ARHGEF3 (ENST00000338458) Intron 2-3	C8orf38 (ENST00000396124) Intron 2-3
GAGAAGGAGTCTTGCTCTGTCACCCAGGCT	TTTATTCTGGAACATTTCCATAGCCTTTCT
LRBA (ENST00000357115) Intron 45-46	NFKB1 (ENST00000394820) Intron 15-16
TAGAGCCTTGCTTTCATGCCCAGTCTGACA	ATTCAAAACATTTTAGGGCCAAATAAAACT
TGGTATTCGAGTAT	
GABPB2 (ENST00000368918) Intron 7-8	BNIPL (ENST00000368931) Intron 1-2
AATGGGAAATAGGGCCGGGCTCAGTGGCTC	TTACTGGGTCAAATCTCGGGTTCCTCAAGG
LAMA5 (ENST00000252999) Intron 11-12	AATK (ENST00000326724) Intron 2-3
TTGCTGCATCGTGGTTAGAATTGGCACACT	GCCTGGCCTCACTAGGCTGCCCCTCTGGAG
TAACCT	
CCNY (ENST00000374706) Intron 3-4	PARD3 (ENST00000374789) Intron 1-2
GTTCTAGGCTGCAGTGAGCCATGATTGTGC	CTCAATTCAACGTATCAGTATATTGTGGGG
ZFPM1(ENST00000319555) Intron 2-3	ZNF469 (ENST00000437464) Exon 2
ATAGCCAAAAATAATAGGACTTTTATACTG	TGTGACGGTGGGCTTCCCGGGAACACCCAC
AATK (ENST00000326724) Intron 2-3	LAMA5 (ENST00000370691) Intron 11-12
GGACGATGTGTGATGCTAGAGCCCAGCCTG	TGAGATTTTCTTTGTTTCTAACCTAGAATG
ITGA10 (ENST00000369304) Intron 23-24	RBM8A (ENST00000330165) 3' downstream
CACTTTGGGAGGCTGAGGCGGGCGGATCAC	GAGAAGACAGAAGTAGAACATGACACAGAC
FCRL1 (ENST00000368176) Intron 1-2	RASAL2 (ENST00000263528) Intron 3-4
CAGTTAAGACCATCTTAAGATGCTCTGTAT	CACGCCTATAATCCCAGCACTTTGGGAGGC
NFKB1 (ENST00000394820) Intron 15-16	FAM160A1 (ENST00000340515) Intron 2-3
GCAGCCTTGGGTAGGCAGTTGGGAGGGACC	ACAGATGCTGTGTGTTAAGGCAGGCTGTCA

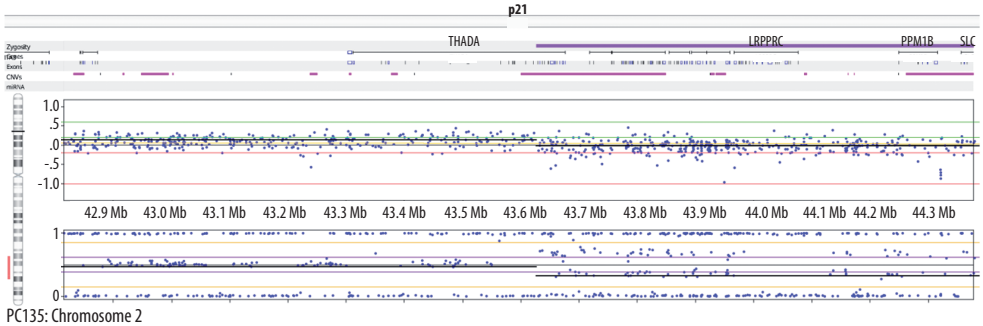
THADA:



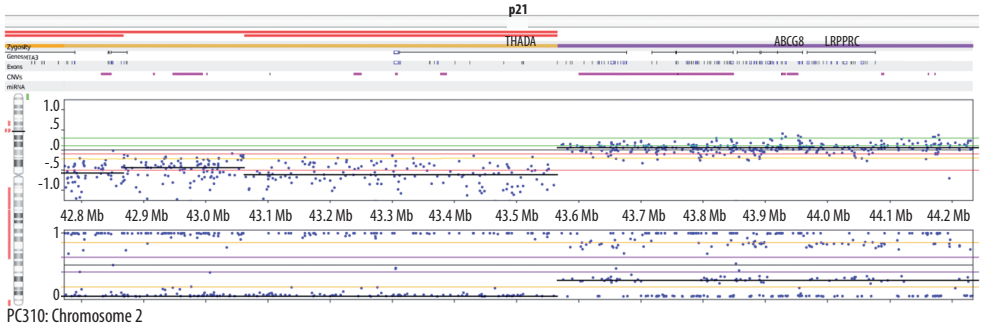
G-321: Chromosome 2



MDA_PCa2B: Chromosome 2

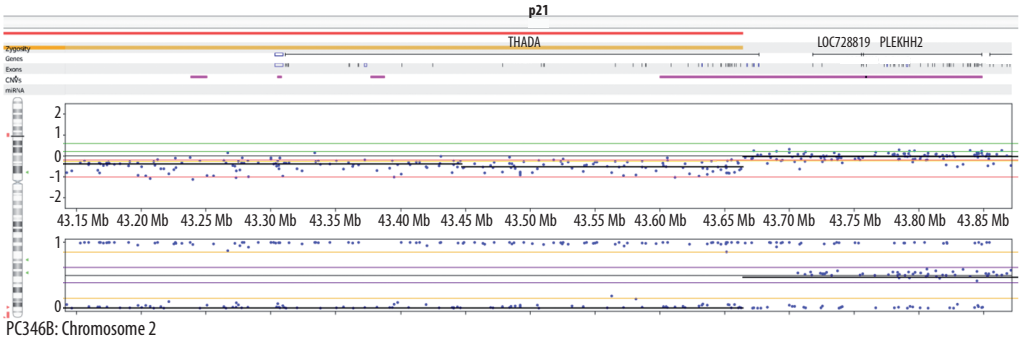


PC135: Chromosome 2

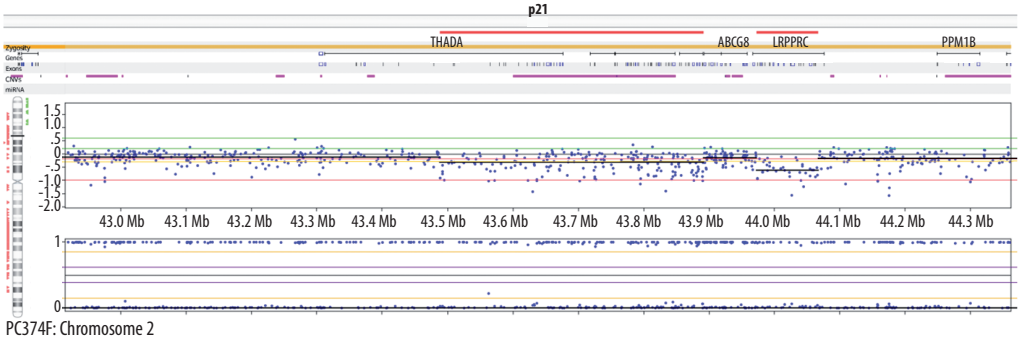


PC310: Chromosome 2

Figure S6: Breakpoints at the SNP level in a dataset of 66 prostate cancer samples

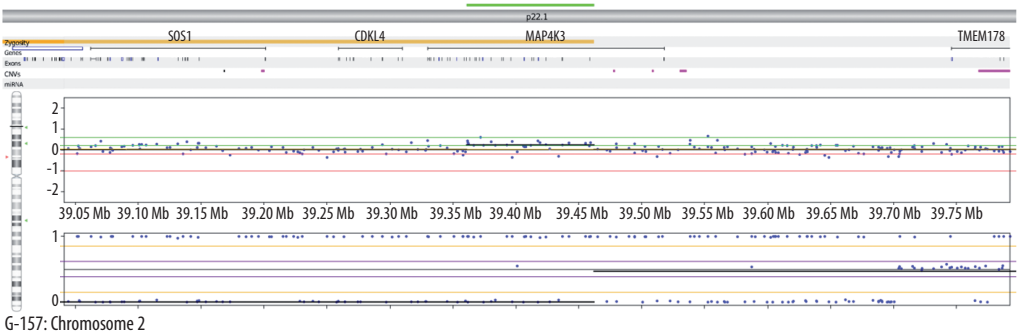


PC346B: Chromosome 2

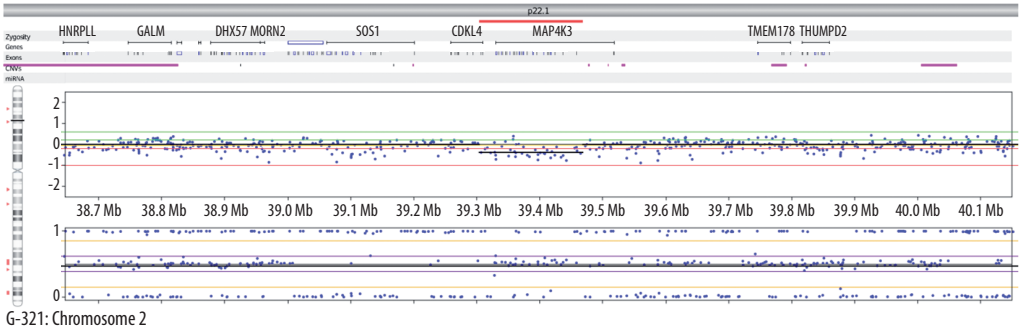


PC374F: Chromosome 2

MAP4K3:

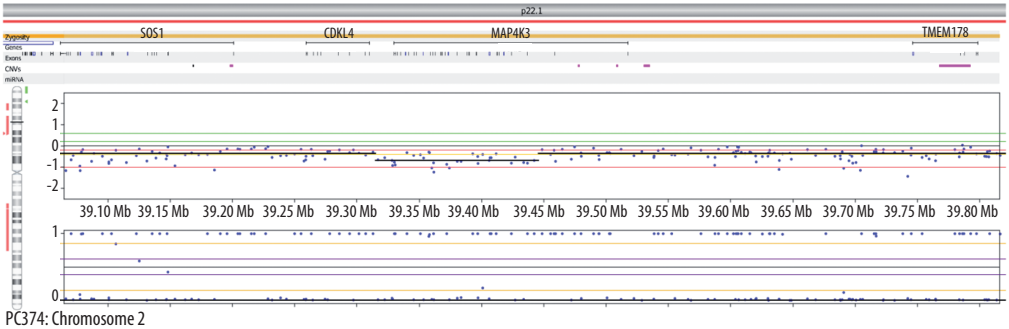


G-157: Chromosome 2

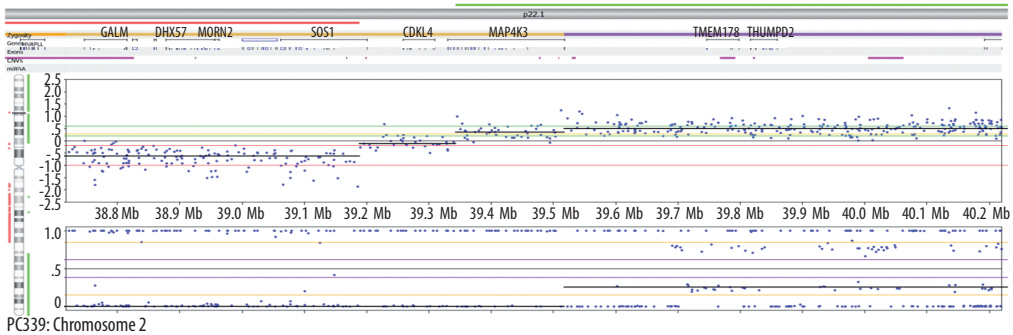


G-321: Chromosome 2

Figure S6: Continued



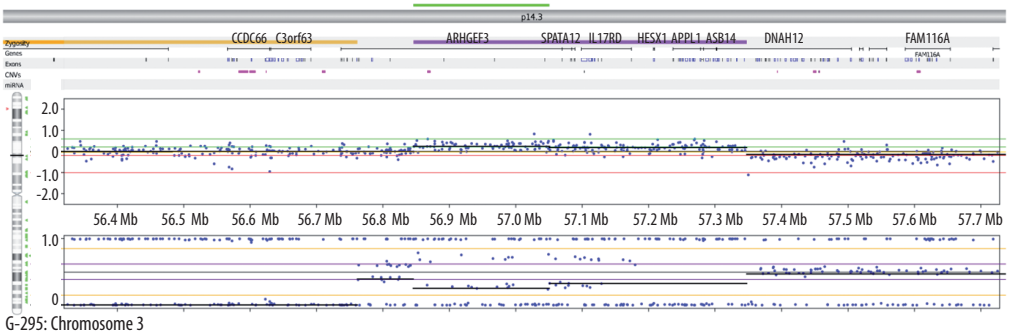
PC374: Chromosome 2



PC339: Chromosome 2

ARHGEF3:

PC135: Chromosome 3



G-295: Chromosome 3

Figure S6: Continued

LRBA:

Figure S6: Continued

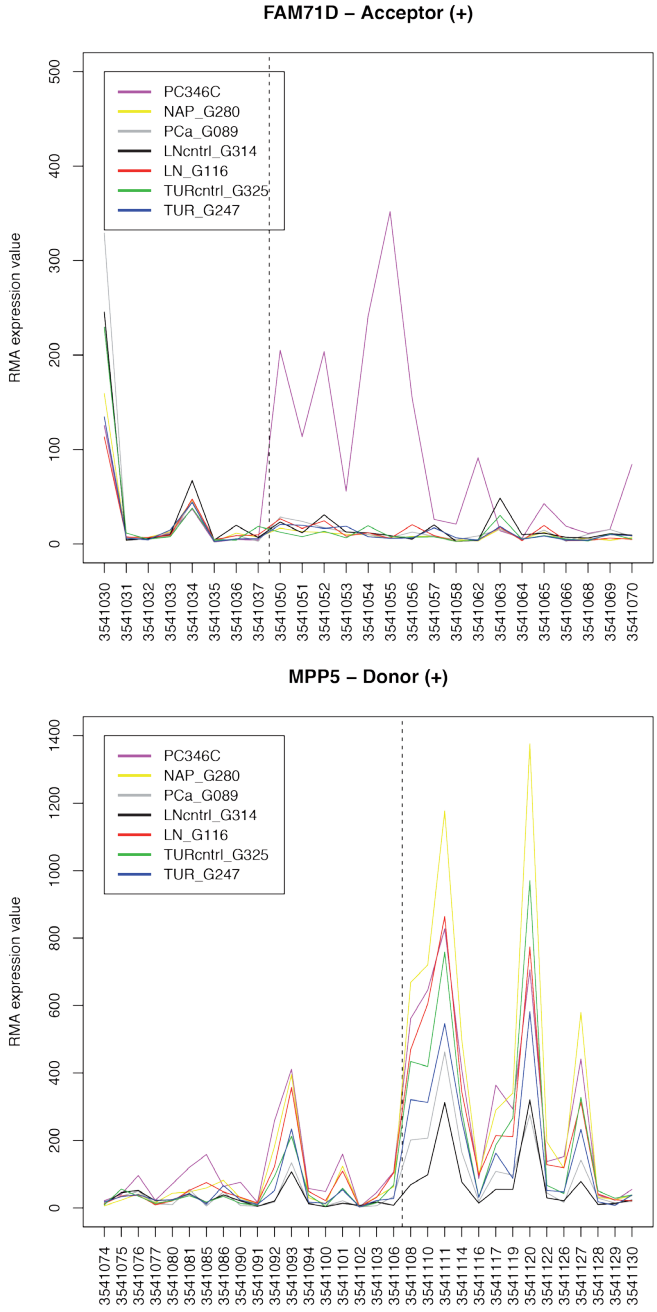


Figure S7: RMA expression of the FAM71D and MPP5 genes in a subset of prostate cancer samples and normal samples is plotted according to the ensemble probeset annotation. The orientation of the genes are given (+ or -) and the dotted line corresponds to the position of the genomic break inside the gene.

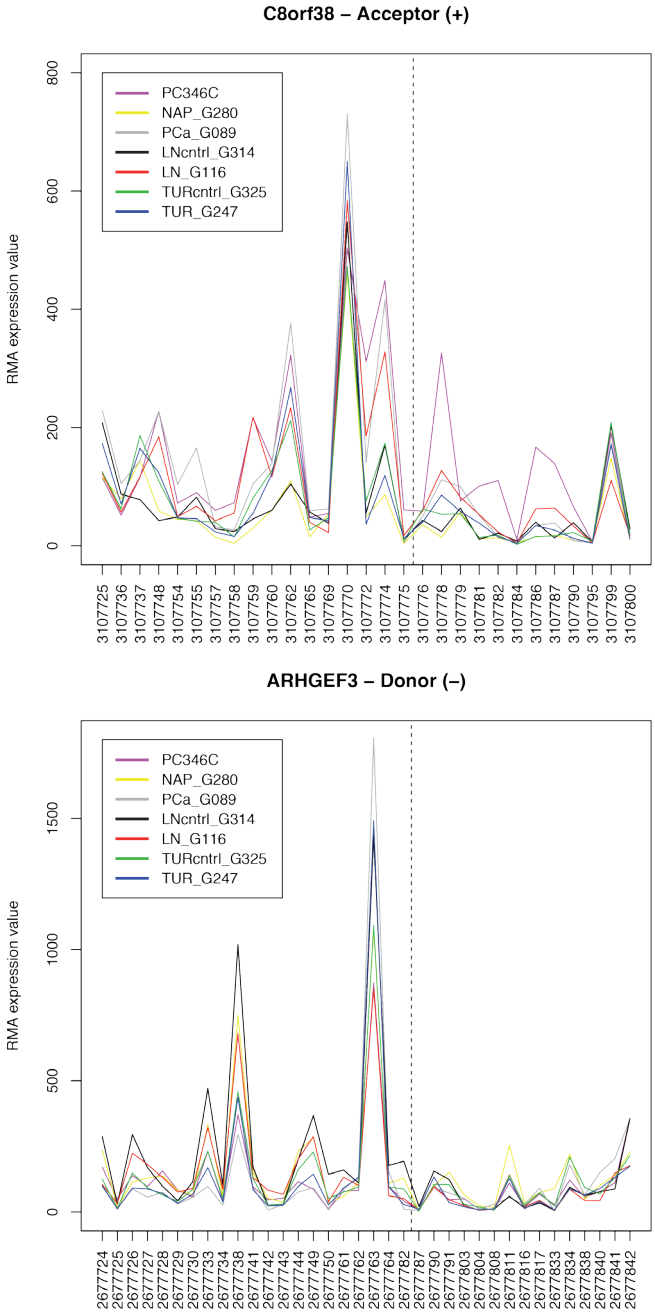


Figure S8: RMA expression of the C8orf38 and ARHGEF3 genes in a subset of prostate cancer samples and normal samples is plotted according to the ensemble probeset annotation. The orientation of the genes are given (+ or -) and the dotted line corresponds to the position of the genomic break inside the gene.

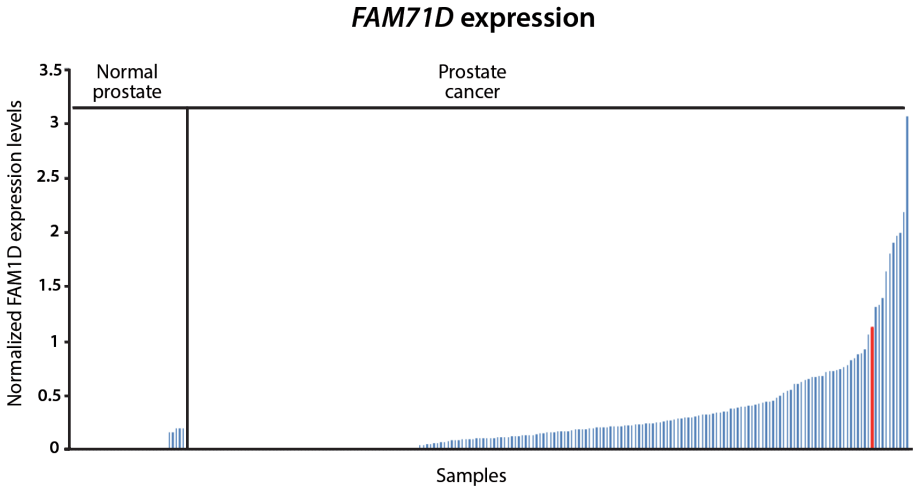


Figure S9: Expression levels of *FAM71D* in a cohort of 201 prostate cancer samples and 33 normal prostate samples. The expression levels were assessed using a custom TaqMan probe against the terminal part of *FAM71D*. This probe was tested for a positive result in the PC346C sample. All expression values were further normalized using the *GAPDH* housekeeping gene. The PC346C sample is depicted in the graph as a red bar. All samples with expression levels above the PC346C sample were considered to have an overexpression of *FAM71D*.

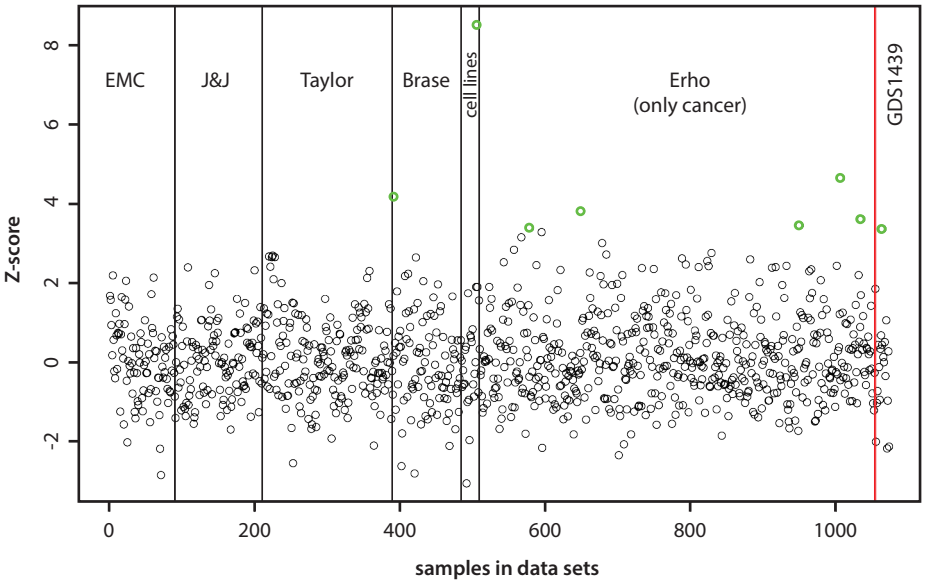


Figure S10: *FAM71D* expression plot among 6 datasets. The expression outliers are depicted in green. References are as follows: EMC – cell lines Boormans JL, Korsten H, Ziel-van der Made AJ, van Leenders GJ, de Vos CV, Jenster G, Trapman J. Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. Int J Cancer. 2013 Jul 15;133(2):335-45; J&J – unpublished; Taylor - GSE21034 Taylor BS et al, Integrative Genomic Profiling of Human Prostate Cancer, Cancer Cell, Volume 18, Issue 1, 11-22, 24 June 2010; Heidelberg - GSE2907 Brase JC et al. TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling. BMC Cancer 2011, 11:507; Erho (Mayo) - GSE46691 Erho N et al. Discovery and Validation of a Prostate Cancer Genomic Classifier that Predicts Early Metastasis Following Radical Prostatectomy. PLoS One. 2013 Jun 24;8(6):e66855 and GDS1439 Varambally S, Yu J, Laxman B, Rhodes DR et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 2005 Nov;8(5):393-406.

A)

Sample	# breaks by SNP arrays	# of breaks confirmed by NGS within: 1000 bp	5000 bp	10,000 bp	50,000 bp	100,000 bp
G089	43	27	30	35	38	38
PC346C	42	22	25	30	33	35

B)

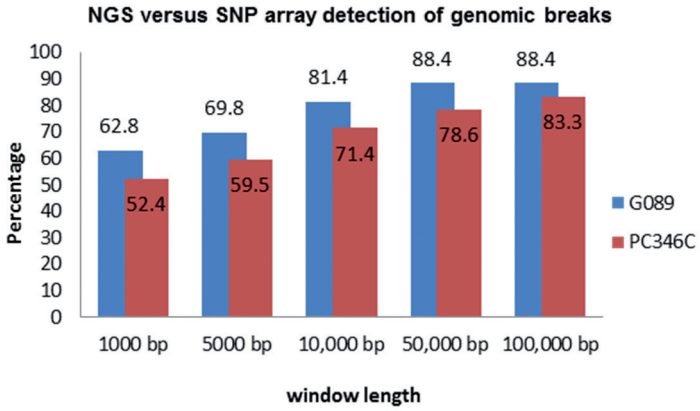


Figure S11: Comparison between the data generated by SNP arrays and NGS. The PC346C and G089 samples were assessed for clear breaks detected by the SNP arrays. This detection was based on the 2-fold difference between the averages of 20 SNP probes on each side of the break. The breaks were manually curated to exclude low coverage SNP areas. The locations of these breaks were then analyzed in the NGS data using a range of window sizes (a). The graph represents the percentage of breaks detected by DNP arrays that were also reported in the NGS data. An increase in the window size (bp) used to search for the breaks detected by the SNP arrays resulted in a higher percentage of overlap between the two sets of data. The minimal overlap percentage is 52.4% whereas the maximal reaches up to 88.4% (b).

The SNP array data of PC346C and G089 samples showed distinct profiles, with more copy number variation events in G089 than in PC346C. This pattern was not concordant with the number of SVs reported for these samples as both had closely the same number of events (Table 1). In order to determine the overlap between these two different sets of data, SNP arrays and next generation sequencing (NGS), we selected the clearest breaks detected by SNP arrays and searched if these were also reported in the NGS data (a). The criteria used to select clear breaks at the SNP level was based on the difference in the average of 20 SNP probes before and after the break. The highest values were then reported and checked manually to exclude low coverage SNP regions such as the ones found in the Y chromosome. We used an increasing window size to search for the genomic locations of the SNP breaks in the NGS dataset. Both PC346C and G089 had an overlap of more than 80% with a window length of 100Kb (b). This significant overlap increased the confidence in the data provided by the NGS.

Table S12: Inventory of validated genes with genomic abnormalities detected by SNP arrays in other prostate cancer samples

Genes	Samples with SNP abnormality
THADA	**PC346B, PC374, PC135, PC310, MDA PCa2b, G321**
MAP4K3	**PC346B, PC374, PC339, G157, G321**
ARHGEF3	**PC135, G295**
LRBA	**PC374, 22RV1**

Supplementary methods

1. Complete Genomics data processing:

Paired-end sequencing for all DNA samples was performed with the Complete Genomics (CG) service provider using a proprietary sequencing-by-ligation technology and primary data analysis, including image analysis, base calling, alignment and variant calling [1]. Reads were mapped to the NCBI Build 36.1 reference genome and mappings were expanded by local de *novo* assembly on all regions of the genome that contain single nucleotide variations (SNVs) relative to the reference genome [1]. Structural variations (SV) reported in the junctions files do not contain SNVs, inserts and deletions (indels) or substitutions (subs) but include information on junctions represented by 3 or more discordant mate pairs. The format of the junctions in the "*allJunctionsBeta*" files [2] was transformed using the table below, which calculates start and end locations of the junctions.

Strand	Left Start	LeftEnd	RightStart	RightEnd
-	Left Position	Left Position + Left Length	Right Position – Right Length	Right Position
+	Left Position – Left Length	Left Position	Right Position	Right Position + Right Length

The junctions are then filtered for normal samples with the union of the all control samples used as a global normal control. Duplicates were filtered out such that if the left and right positions and chromosomes of one junction exactly matched that of another, only one of the junctions was kept. The two prostate cancer samples were filtered using the global normal control by removing any junction in the sample that had a matched normal junction. A matched normal junction has a left/right position within 10 base pairs of the left and right positions in the sample junction.

The remaining junctions (left and right positions) are annotated with gene features (exons, gene symbol) derived from UCSC genome browser [3] using the rtracklayer package [4] in R. The junctions located inside such regions are further categorized as junctions with different genes on both sides, with the same genes or junctions with only one gene on either the left or right side. The following table depicts how the single intrachromosomal structural variation events are called per junction. Note from personal communication with CGI: "+/+ junction may indicate a deletion or insertion, but our current process only detects deletions."

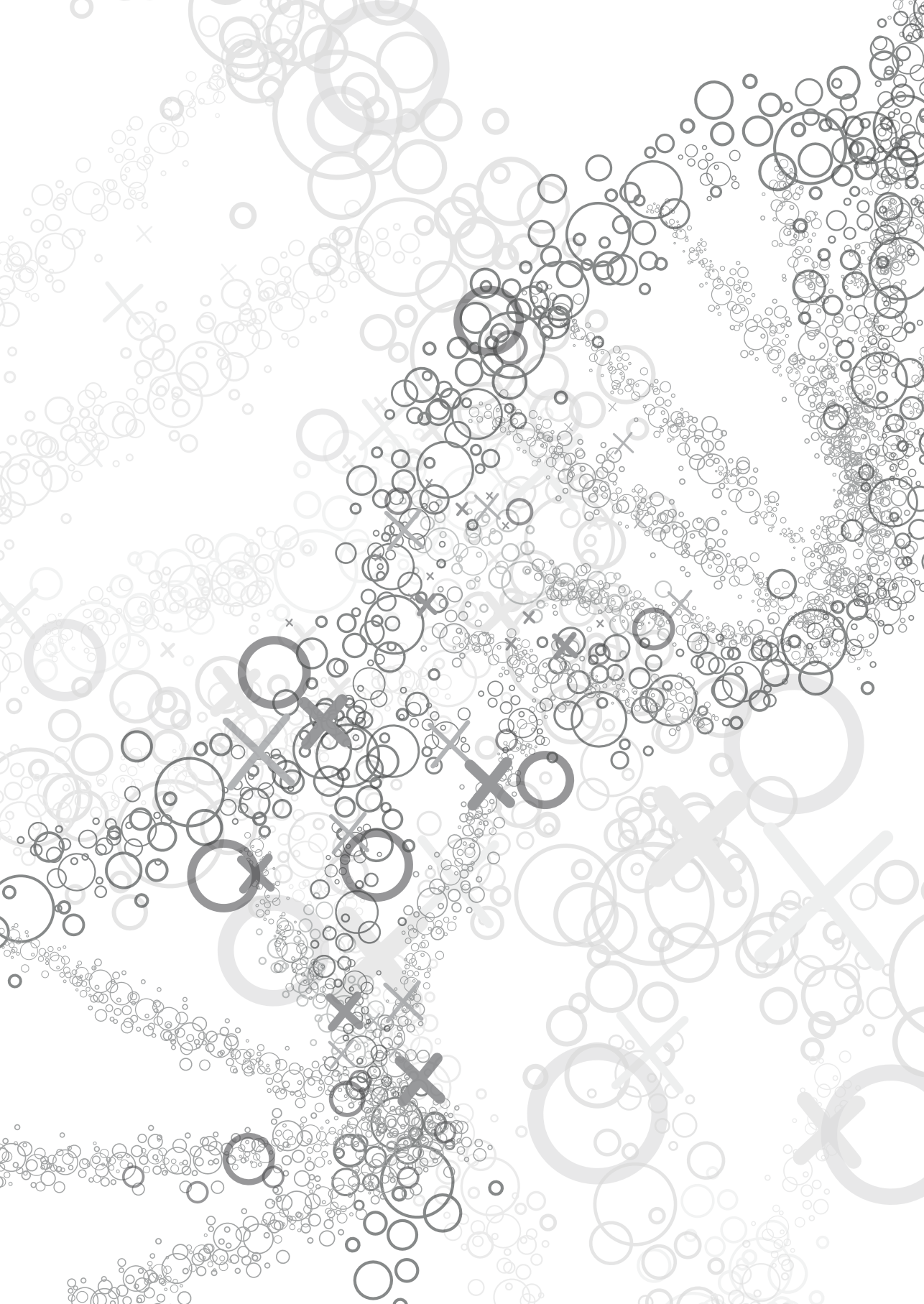
Junction LeftStrand	Junction Right Strand	Single Event
+	+	deletion
-	-	inversion or deletion
+	-	inversion
-	+	inversion

2. CNV analysis of CGI prostate cancer samples data:

The algorithm CNV-Seq was used to determine copy number variation from the CGI data. This algorithm uses a sliding window analysis method with a window size of 2000bp and a slide of 1000bp. CNV-Seq determines whether the coverage in each region is significantly different between test and reference (reference is hg18). Four or more consecutive regions with log2 relative coverage (essentially test coverage / reference coverage) greater than 0.6 classifies as a gain and lower than -0.6 classifies as a loss. The "output".csv file for each chromosome in each sample contains the called ploidy which is determined as specific below:

0 if cnv_log2 < -1.35
1 if -1.35 ≤ cnv_log2 < 0
2 if not in region of loss or gain
3 if 0 ≤ cnv_log2 < 1.35
4 if cnv_log2 ≥ 1.35

1. Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V, Martin B, Ballinger DG, Drmanac R. **Computational techniques for human genome resequencing using mated gapped reads.** *J. Comput. Biol.* 2012, 3:279-292.
2. http://www.bx.psu.edu/~giardine/tests/tmp/DataFileFormats112.pdf
3. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. The human genome browser at UCSC. *Genome Res.* 2002, **12(6)**:996-1006.
4. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009, **25(14)**:1841-2.



Chapter 4

A mononucleotide repeat in PRRT2 is an important, frequent target of mismatch repair deficiency in cancer

Inês Teles Alves [1,2]
David Cano [2]
René Böttcher [1]
Martin van Royen [2]
Hetty van der Korput [2]
Winand Dinjens [2]
Guido Jenster [1]
Jan Trapman [2]

Departments of Urology[1], Pathology[2]
Erasmus MC, Rotterdam, the Netherlands

Manuscript submitted

Abstract

The DNA mismatch repair (MMR) system contributes to maintaining genomic stability by the correction of DNA replication errors. MMR deficiency increases mutation frequency and predisposes to cancer. Whole genome sequencing data provides a complete overview of novel important targets in MMR deficiency. We have performed whole genome sequencing of the MMR deficient PC346C prostate cancer cell line. Overall, we detected 1196 mutations in PC346C. Comparison with DNA from a MMR proficient prostate tumor showed that frameshifts in mononucleotide repeat sequences were most specific for MMR deficiency. A selection of genes with mononucleotide frameshift mutations was further analyzed in a panel of prostate, ovarian, endometrial and colorectal cancer cell lines. *PRRT2* and *DAB2IP* were identified as novel genes frequently mutated in cell lines and also in genomic DNA from MMR deficient colorectal and endometrial cancer patient samples. Functional studies showed that the mutated form of *PRRT2* (ΔPRRT2) promoted both cellular proliferation and migration.

Introduction

The DNA mismatch repair (MMR) system is highly conserved and contributes to maintain genomic stability through the correction of mismatched base pairs occurring during the DNA replication process (1). Defects in this repair mechanism are associated with mutations in microsatellite repeat sequences (2), which consist of tandemly repeated motifs of one (mono) to six (hexa) nucleotides (3). The repetitive nature of these sequences makes them highly prone to insertion/deletion mutations caused by DNA polymerase slippage thereby altering the length of the microsatellite (4). The mismatch repair heterodimers MSH2/MSH6 and MSH2/MSH3 detect replication errors and recruit the MLH1/PMS2 complex which in turn degrades and resynthesizes the mutated stretch (5). In case the MMR system is defective, these errors will not be corrected and will lead to frequent microsatellite instability (MSI) (2).

The majority of microsatellite sequences are located in non-coding DNA with limited consequences to cell functions (6). But defects in the MMR system also give rise to a mutator phenotype with an increased rate of frameshift mutations in genes with coding microsatellites and can therefore increase cancer predisposition (7). Most of the gene mutations associated with MSI cancers have been observed in mononucleotide repeats (MNR). Affected genes are involved in signal transduction pathways (*TGFβRII, IGFIIR* and *PTEN*) (8-10), apoptosis (*BAX* and *CASP5*) (11, 12), DNA repair (*MBD4*) (13), and transcriptional regulation (*TCF-4, EPHB2, AXIN2*) (14-16). The proteins resulting from these frameshift mutations can contribute to tumor progression by functional inactivation, a dominant negative effect, or a gain of function (17-19).

Germline mutations in MMR genes are present in over 90% of Lynch syndrome patients, which are characterized by MSI and an increased risk of colorectal cancer and several other cancer types (20). The MSI phenotype is also observed in approximately 15% of sporadic colorectal cancers (21), 12% of ovarian cancers (22), 20% of endometrial cancers (23, 24), 30% of gastric cancers (25) and less frequent in other cancer types (26). MMR deficiency is also frequent in *in vitro* growing prostate cancer cell lines. LNCaP cells do not express MSH2 due to gene deletion and *MLH1* is mutated in DU145, resulting in expression of an unstable truncated *MLH1* protein (27, 28). The frequency of MSI reported in prostate cancer patients varies considerably between different studies but it is conclusively lower than in other sporadic cancer types (29, 30).

So far, no whole genome sequencing study has focused on identifying novel genes commonly targeted by MSI in prostate cancer cell lines and patient samples. We performed whole genome paired-end sequencing of two prostate cancer samples, one with a functional MMR system (G089, a late stage prostate cancer patient) and the other with deficiency of the MMR system due to absence of MSH2 expression caused by point mutations in both gene alleles (PC346C, a prostate cancer cell line) (31).

Numerous mutations were observed in G089 and PC346C. Mutations in mononucleotide repeats were most specific for the MMR deficient PC346C sample. A potentially

interesting set of 17 PC346C-specific genes with MMR deficiency-associated mutations (insertions/deletions in mononucleotide repeats) was defined and further evaluated in a larger panel of prostate, colorectal, endometrial and ovarian cancer cell lines. We identified proline-rich transmembrane protein 2 (*PRRT2*) and DAB2 interacting protein (*DAB2IP*) to be frequently mutated in all different cancer cell line types. Further analysis showed that both genes were also frequently mutated in colorectal and endometrial cancer patient samples. Functional studies revealed PRRT2 to be implicated in cellular proliferation and migration with the truncated MSI-derived PRRT2 form promoting both processes.

Results

Overview of microsatellite mutations in gene sequences

To identify novel gene mutations caused by MMR deficiency in prostate cancer, we analyzed the whole genome sequence of the MMR-deficient prostate cancer cell line PC346C and the MMR-proficient prostate cancer patient sample (G089). We selected for mutations that would likely disrupt the normal protein function (insertion/deletion introducing frameshifts, and mutations in the start or stop codon). The sequence flanking each gene mutation (10 nucleotides up and downstream) was retrieved using the UCSC genome browser. We selected for mutations occurring in microsatellite sequences present in the PC346C sample and absent in G089, thereby enriching for mutations that are caused by MMR deficiency. First, the type and length of repeats were used to attain an overview of the mutation spectrum in both prostate cancer samples (Figures 1A, 1B and 1C).

We included in our studies repeats with three or more repeat units, depending on the repeat composition (mono-, di- or three-nucleotide repeat). Overall, we detected 1196 gene mutations in PC346C and 760 gene mutations in G089 (Tables S1 and S2, see supplementary data. From the 1196 gene mutations in PC346C, 303 were in mononucleotide repeats (MNRs), 79 in dinucleotide repeats (DNRs) and 19 in trinucleotide repeats (TNRs) (6x n_1 for MNR, 3x n_2 for DNR and 3 x n_3 for TNR) (Figure S1A). In G089, 47 gene mutations were in MNRs, 56 in DNRs and 19 in TNRs (6x n_1 for MNR, 3x n_2 for DNR and 3 x n_3 for TNR) (Figure S1B).

PC346C had a 1.5-fold increase in the number of mutated genes as compared to G089. The number of mutations found in MNR sequences of 6 nucleotides was 6-fold higher in PC346C. A 6-fold increase in MNR gene mutations in PC346C was also observed for MNRs of length n8. Since MNRs showed, compared to di- and tri-nucleotide repeats, a substantial higher incidence of mutations in MSI+ cancers we focused on these repeats.

We separated the mutated genes in PC346C and G089 in A, T, C and G MNRs categories of different lengths (Table S3). The UCSC build hg18 was used as reference for gene annotation. The number of C and G MNRs in reference genes was considerable lower

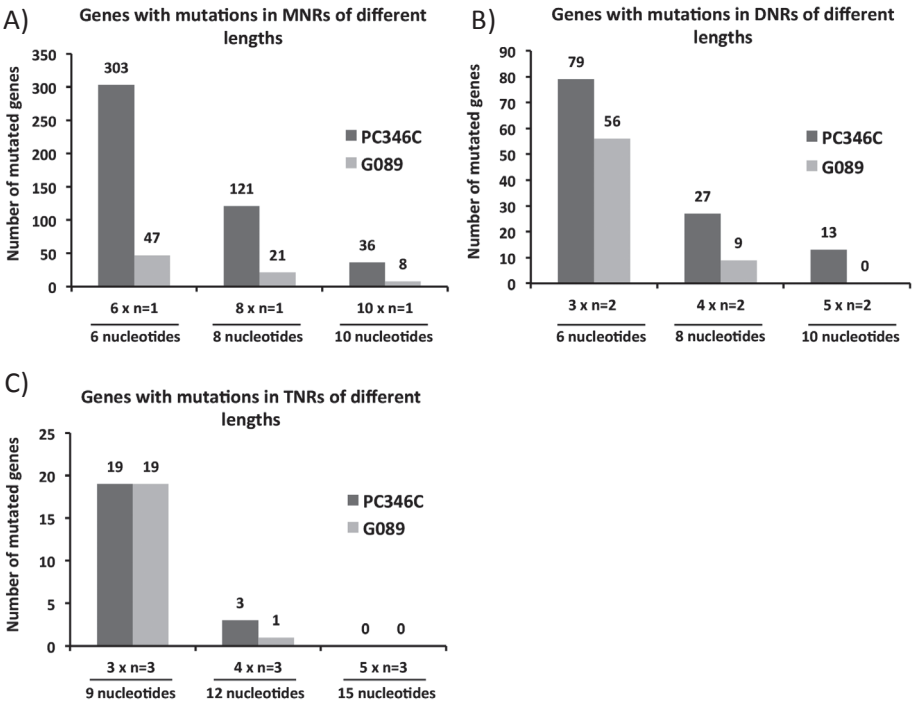


Figure 1: Overview of the mutation spectrum in G089 and PC346C genomic DNA. Genes with mutations in mononucleotide repeats (MNRs) (A), dinucleotide repeats (DNRs) (B) and trinucleotide repeats (TNRs) (C) were distributed according to the length of the mutated repeat sequence. The repeat unit length for MNRs is one nucleotide (n=1), for DNRs it is two nucleotides (n=2) and for TNRs it is three nucleotides (n=3). The x-axis shows the number of repeat units: 6, 8 and 10 in case of MNRs; 3x, 4x and 5x in case of DNRs and TNRs.

than that of A and T repeats. As expected the mutations in C and G MNRs in PC346C and G089 was also much lower than that of A and T MNRs (Table S3).

Novel target genes of microsatellite instability

The comparison between the whole genome sequence of the MMR-deficient prostate cancer cell line PC346C and the MMR-proficient prostate cancer sample G089 allowed the identification of novel candidate target genes of MMR deficiency. Our selection procedure aimed to identify novel gene mutations associated with MMR deficiency with expected functional impact. First, we filtered for mutations in MNR sequences in PC346C, which were 6 nucleotides or longer. We excluded LOC annotated gene symbols and genes already known to be associated with MMR deficiency or MSI (text mining software Anni 2.0, http://biosemantics.org/anni/). We identified over 35 mutated genes in PC346C, which were already known as targets of MSI or are associated with MMR deficiency including *TGFβRII, PTEN, BAX, MLH3, MSH6* and *MSH2*. The frequent

identification of known MMR-associated genes confirmed the validity of our approach. This procedure left 119 genes for further selection. Next, Ingenuity Pathway Analysis (IPA) was used to identify genes with functions most likely to be cancer-associated, like cell cycle, cellular growth, apoptosis and other important cell signaling pathways. Finally, we checked the expression profile of candidate genes using Affymetrix exon array data of 90 prostate cancer RNA samples and 17 normal controls (supplementary materials and methods). A final subset of 14 novel genes showing MNR mutations and differential expression between normal and prostate cancer samples was assessed for their mutation status in a larger panel of cancer cell lines (Table 1). For comparison we added three known MSI target genes (*PHACTR4, MLL3* and *CEP164*).

Table 1: List of microsatellite instability target genes in PC346C. Genes are ordered first by the frequency of mutation in MMR deficient prostate cancer samples and second alphabetically. The 3 MMR associated genes used as a positive control are depicted in red.

Gene	Name	Length of repeat	Mutation in PC346C
CNOT1	CCR4-NOT transcription complex, sub 1	A13	del AA
DAB2IP	DAB2 interacting protein	C8	del C
PRRT2	proline-rich transmembrane protein 2	C9	ins CC
TTC3	tetratricopeptide repeat domain 3	A8	del A
CEP164	centrosomal protein 164kDa	A11	del A
PHACTR4	phosphatase and actin regulator 4	A10	del A
TROVE2	TROVE domain family, member 2	A9	del A
ZFR	zinc finger RNA binding protein	T9	del T
MLL3	myeloid/lymphoid or mixed-lineage leukemia 3	T9	del T
ANLN	anillin, actin binding protein	A7	del A
ANUBL1	AN1, ubiquitin-like, homolog	A7	del A
EPRS	glutamyl-prolyl-tRNA synthetase	T6	ins T
KCNMA1	potassium large conductance calcium-activated channel, subfam M, α1	T9	del T
PDS5A	PDS5, regulator of cohesion maintenance, homolog A (S. cerevisiae)	A6	del A
SFRS12IP1	SFRS12-interacting protein	T10	ins T
TSHZ2	teashirt zinc finger homeobox 2	C7	del C
USP42	ubiquitin specific peptidase 42	A8	del A

Frequency of microsatellite instability target genes

Microsatellite status of the set of target genes was examined in three MMR-proficient prostate cancer cell lines and in four MMR-deficient prostate cancer cell lines (Figure 2, Table S4, Figure S2). None of the MMR-proficient prostate cancer cell lines showed MSI in the selected gene panel. Controls *CEP164, PHACTR4* and *MLL3* were mutated in two or more of the MMR-deficient cell lines. We found that *CNOT1, DAB2IP* and *PRRT2* were mutated in at least three MMR-deficient cell lines (Figure 2, Table S4).

Next, we expanded the cell line panel with ten colorectal, four endometrial and three ovarian cancer MMR-deficient cell lines and control MMR-proficient cell lines of each cancer type. As observed in the MMR-proficient prostate cancer cell lines, colorectal, endometrial and ovarian cancer MMR-proficient cell lines showed no mutations in the analyzed genes (Table 2 and Table S5). The control genes *CEP164, PHACTR4* and *MLL3* displayed as expected high mutation frequencies, ranging from 30% to 80% in the MSI cancer cell lines (Table 2 and Table S5).

Also here we identified *CNOT1, DAB2IP* and *PRRT2* as novel MSI target genes, in the MMR-deficient cancer cell lines, although with varying frequency. The genes displayed mostly deletions and to a lesser extend insertions (Table 2 and Table S5). The mutations shifted the open reading frame of the affected genes and consequently predicted to result in synthesis of truncated proteins due to premature termination (Table S1). *CNOT1* contains a n_{13} A repeat in the open reading frame located 10 nucleotides upstream of the stop codon. A frameshift mutation at this position might not be critical for the protein and no further research was conducted for this gene. The n_9 C repeat in *PRRT2* and the n_8 C repeat in *DAB2IP* displayed a mutation pattern reflecting both insertions and deletions. (Figure S3A, S3B).

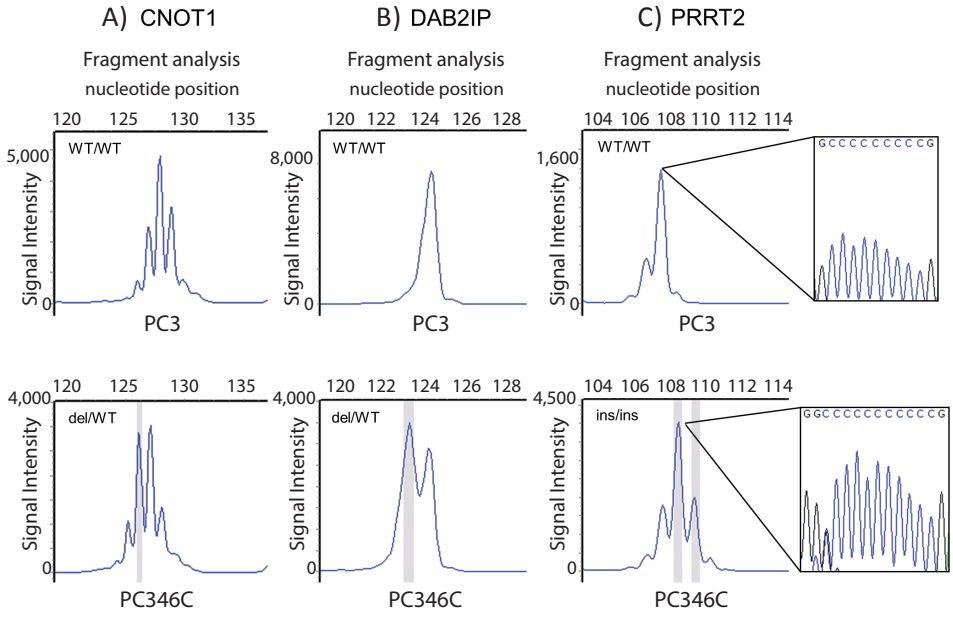


Figure 2: MNR repeat analysis of top mutated genes in prostate cancer in genomic DNA from prostate cancer cell lines PC346C (MMR deficient) and PC3 (MMR proficient). Fragment size analyses of PCR amplified fragments of CNOT1 (A), DAB2IP (B) and PRRT2 (C) containing the repeat are presented. The highlighted peaks represent a mutant allele. PCR primers are presented in supplementary data. WT: wild type; del: deletion; ins: insertion. Indicated PRRT2 PCR fragments are flanked by sequence analysis of the repeat.

Table 2: Number of MSI positive (MSI) and negative (MSS) prostate, colorectal, endometrial and ovarian cancer cell lines with mutations in the set of MSI target genes in PC346C. The 3 control genes are depicted in red. Three MSS cell lines were available for prostate, colorectal and endometrial cancer and one MSS cell line for ovarian cancer. The number of MSI cell lines available for prostate, colorectal, endometrial and ovarian cancer were four, ten, four and three respectively.

	Prostate Cancer		Colorectal Cancer		Endometrial Cancer		Ovarian Cancer	
	MSS	MSI	MSS	MSI	MSS	MSI	MSS	MSI
CNOT1	0/3	4/4	0/3	9/10	0/3	4/4	0/2	1/3
DAB2IP	0/3	3/4	0/3	6/10	0/3	1/4	0/2	0/3
PRRT2	0/3	3/4	0/3	4/10	0/3	4/4	0/2	2/3
TTC3	0/3	2/4	0/3	5/10	0/3	1/4	0/2	0/3
CEP164	0/3	3/4	0/3	8/10	0/3	3/4	0/2	2/3
PHACTR4	0/3	3/4	0/3	7/10	0/3	3/4	0/2	1/3
TROVE2	0/3	1/4	0/3	5/10	0/3	4/4	0/2	1/3
ZFR	0/3	2/4	0/3	4/10	0/3	3/4	0/2	1/3
MLL3	0/3	2/4	0/3	8/10	0/3	3/4	0/2	1/3
ANLN	0/3	1/4	0/3	0/10	0/3	0/4	0/2	0/3
ANUBL1	0/3	1/4	0/3	2/10	0/3	0/4	0/2	0/3
EPRS	0/3	1/4	0/3	0/10	0/3	0/4	0/2	0/3
KCNMA1	0/3	1/4	0/3	6/10	0/3	1/4	0/2	1/3
PDS5A	0/3	1/4	0/3	0/10	0/3	1/4	0/2	0/3
SFRS12IP1	0/3	1/4	0/3	6/10	0/3	2/4	0/2	1/3
TSHZ2	0/3	1/4	0/3	10/10	0/3	3/4	0/2	3/3
USP42	0/3	1/4	0/3	4/10	0/3	4/4	0/2	0/3

Frequency of *PRRT2* and *DA2IP* repeat mutations in MSI patient cancer samples

Next, we checked whether the mutation frequencies of *PRRT2* and *DAB2IP* observed in the cancer cell line panels could be confirmed in primary cancer samples. We tested a total of 80 prostate cancer patient samples, including late stage and metastasized cancers. None of the samples displayed the mutations in either of the genes (data not shown), not unexpected because MSI is not frequent in clinical prostate cancer. Further, we tested 24 colorectal and 24 endometrial MSI patient cancer samples. We found *PRRT2* to be mutated in 15 out of 24 colorectal cancer patients and in 11 out of 24 endometrial cancer patients (Figure S4, Table S6). Although the mutation frequency in the *DAB2IP* MNR was similar (11/23), only one endometrium cancer sample displayed a mutation in the repeat. To further evaluate the significance of the mutation frequency of *PRRT2* and *DAB2IP* in clinical MSI colorectal and endometrial cancer patient samples, we compared their frequency to the average mutation frequency for common MSI target genes published in databases (SelTarbase (33)) (Table S7). Compared to other genes, *PRRT2* displayed a very high mutation frequency in both colorectal and endometrial cancer and was selected for functional studies.

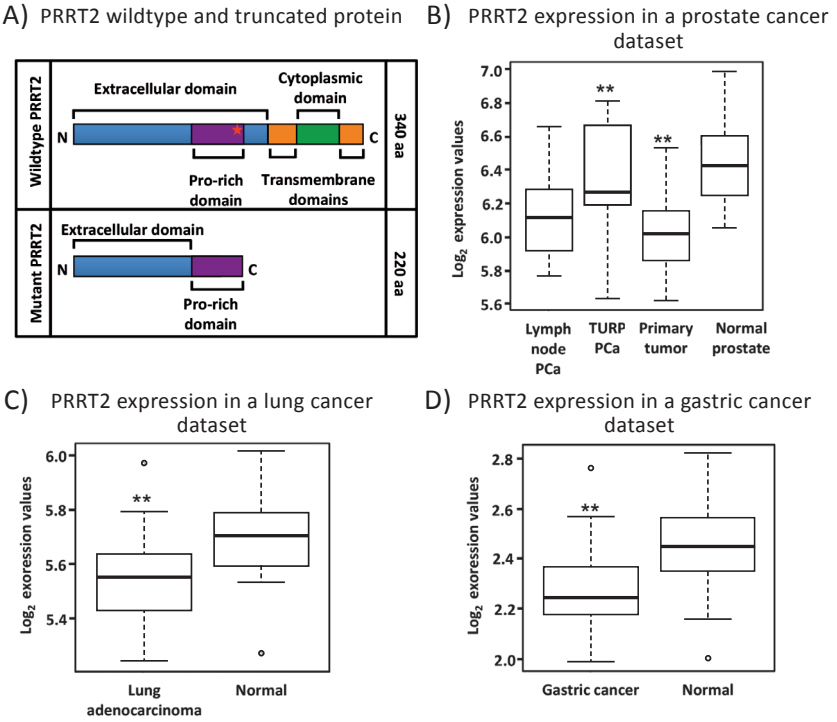


Figure 3: Expression pattern of *PRRT2* mRNA. (A) Schematic representation of the wildtype and the mutant *PRRT2*. The insertion of C and CC in the 9C repeat of PC346C leads to a truncated protein lacking the predicted transmembrane and cytoplasmatic domains. The frameshift occurs within the proline-rich region of *PRRT2*. (B) Box plot analysis of *PRRT2* mRNA in prostate cancer lymph node metastasis (LN-PCa, 12 samples), transurethral resections (TURP- PCa, 10 samples), primary prostate cancer (PCa, 56 samples) and normal adjacent prostate tissue (NAP, 12 samples). (C) Expression levels of *PRRT2* in lung adenocarcinoma (22 samples) and normal lung tissue (22 samples). (D) Expression levels of *PRRT2* in gastric cancer (20 samples) and normal gastric tissue (20 samples). Wilcoxon testing was used in (B), (C) and (D) to determine significant differences in expression levels.

The *PRRT2* gene

PRRT2 encodes a 340 amino acids protein that has two predicted transmembrane domains. This protein is poorly characterized, although recently a link between *PRRT2* mutations and a rare neurological disease, paroxysmal kinesigenic dyskinesia, was established (34). Whether *PRRT2* has a role in cancer is unknown. Frameshifts in the C-repeat would result in expression of truncated proteins of 222 or 227 amino acids that lack the predicted transmembrane domains (Figure 3A). The PC346C prostate cancer cell line is the only cell line we identified, which has a mutation in both *PRRT2* alleles. To investigate a possible role of *PRRT2* in cancer we first documented *PRRT2* expression levels by RNA sequencing in available cancer data sets (Supplementary materials and

methods). Overall, *PRRT2* showed a consistent decrease in expression in cancer as compared to normal in prostate, lung and gastric tissue samples (Figures 3B, 3C, 3D). Both lung and gastric cancers have been associated with MSI. In order to exclude non-mediated mRNA decay of the mutated *PRRT2* transcript we investigated the expression level of *PRRT2* in PC346C. The elevated expression level detected suggested that the stability of the *PRRT2* transcript is not affected by the mutation (Figure S5).

Biological activity of PRRT2

To assess whether the truncated *PRRT2* (ΔPRRT2) influenced cell proliferation, we stable transfected the normal prostate epithelial cell line PNT2C2 with lentiviral constructs expressing wild type *PRRT2* ($PRRT2^{wt}$) or ΔPRRT2. Efficiency of transfection was assessed by the expression of GFP, independently transcribed under the CMV promoter. Cell proliferation was measured during 4 days after plating of equal numbers of control (GFP only), $PRRT2^{wt}$ and ΔPRRT2 transfected cells.

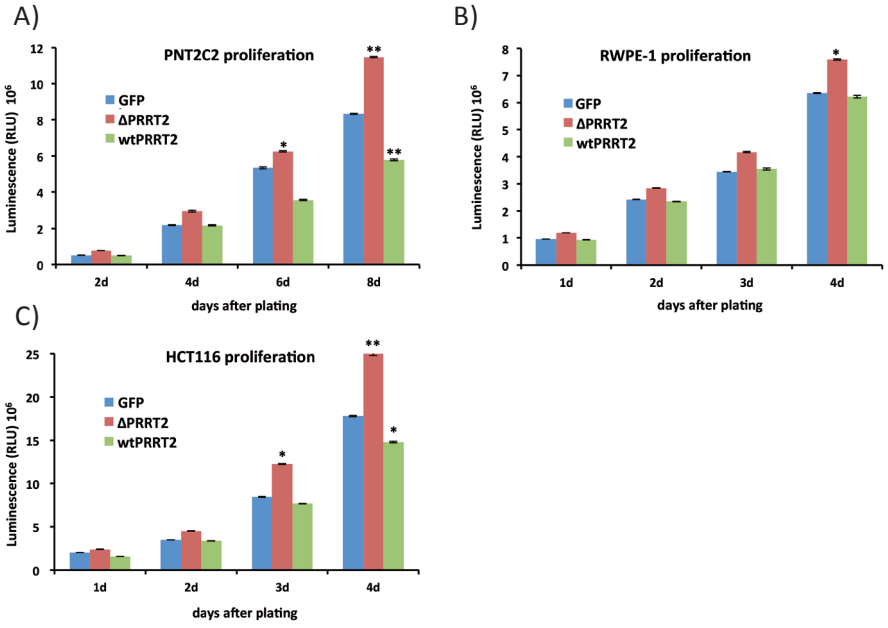


Figure 4: PRRT2 affects cellular proliferation. (A) PNT2C2, (B) RWPE-1 and (C) HCT116 cells were stably transduced with lentiviral vectors expressing the truncated form of PRRT2 (ΔPRRT2), the wildtype PRRT2 (wtPRRT2) and the control GFP vector. Cellular proliferation was measured 1, 2, 3 and 4 days after plating of the RWPE-1 and HCT116 cells, and 2, 4, 6 and 8 days after plating of the PNT2C2 cells. Cellular viability was measured by the reduction of luciferin into ocyluciferin in the presence of ATP. The y-axis displays the RLU (relative light units). Asterisks indicate *P≤0.02 and **P≤0.002, respectively, in a two-tailed Student's t-test.

A significant increase in proliferation capacity of PNT2C2 cells expressing ΔPRRT2 (37.5%) was observed as compared to the control cells at day 8. Conversely, the growth of PNT2C2 cells expressing PRRT2^{wt} was impaired especially as compared to the PNT2C2 cells expressing ΔPRRT2 (Figure 4A).

The ΔPRRT2 effect on cellular proliferation was further validated in the RWPE-1 prostate epithelial cell line and the HCT116 colorectal cancer cell line. Both cell lines displayed the same growth promoting effect of ΔPRRT2. RWPE-1 cells stably expressing ΔPRRT2 showed an increase of 20% in proliferation capacity whereas PRRT2^{wt} did not affect proliferation (Figure 4B). HCT116 cells expressing ΔPRRT2 showed an increase of 41% in proliferation capacity as compared to the control. Similar to PNT2C2, cells expressing PRRT2wt showed a decrease in cellular proliferation (Figure 4C).

The influence of *PRRT2* expression on apoptosis was addressed in a caspase 3/7 activation assay in all ΔPRRT2 and PRRT2^{wt} expressing cell lines described above. In none of the cases we observed a clear difference in apoptosis levels between ΔPRRT2, PRRT2^{wt} and control GFP expressing cells (data not shown).

Finally we addressed the effect of *PRRT2* on cell migration. PNT2C2, RWPE-1 and HCT116 expressing ΔPRRT2, PRRT2^{wt} and GFP were cultured in migration chambers for 24h in both presence and absence of 10% FCS. We detected a significant increase in migratory capacity of ΔPRRT2 expressing cells as compared to the GFP controls (Figure 5A and 5B). Migration of PRRT2^{wt} expressing cells was consistently lower than the GFP control cells, but in all cases, not significantly different.

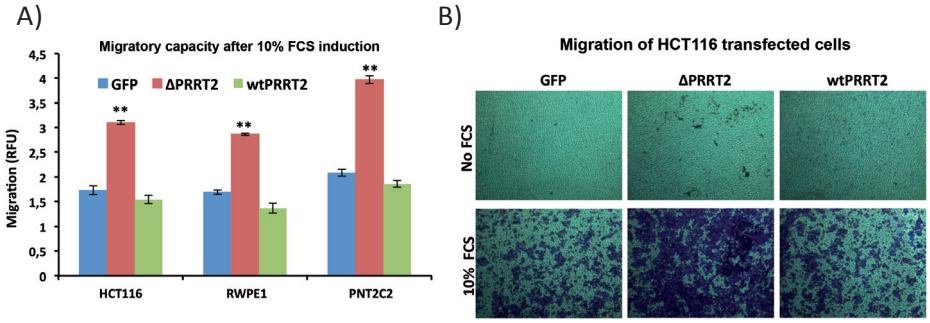


Figure 5: PRRT2 plays a role in cellular migration. (A) Migratory capacity of HCT116, PNT2C2 and RWPE-1 cells was assayed by measuring the cells migrating through a 5 µm polycarbonate membrane. Migration was induced by 10% FCS. Cells were assayed using CyQuant® GR fluorescent dye. All samples were normalized to the fluorescent detection in the absence of chemo attractant (10% FCS). RFU represents the relative fluorescent units. **P≤0.002 was calculated using a two-tailed Student's t-test. (B) Representative illustrations of migration of HCT116 cells expressing ΔPRRT2, wtPRRT2 or control GFP in the presence and absence of 10% FCS (see also legend to Fig. 4).

Discussion

Microsatellite instability in cancer

The most prominent characteristic allowing cancer cells to survive, proliferate and disseminate is the development of genomic instability (35). DNA repair mechanisms are the cell's surveillance systems and defects in this machinery are responsible for accelerating the rate at which cells can accumulate favorable genotypes. The MMR system eliminates base-base mismatches and insertion-deletion loops that arise during DNA replication due to slippage of the DNA polymerase (1). Simple repetitive DNA sequences known as microsatellites are particularly prone to insertions and deletions due to defective MMR (36). As a result, cells with a defective MMR system might have mutation rates 100-1000 fold higher than normal cells (37). Several studies have shown that instability of microsatellites is present in PCa (38, 39). The frequency of microsatellite instability detected in PCa is variable between different studies and ranges from 8% (40) to 35% (41, 42). To elucidate the role of MMR in PCa we have sequenced the complete genome of the MMR-deficient PC346C cell line. This cell line was derived from a transurethral resection of a primary tumor and lacks a functional MSH2 protein due to point mutations in both alleles (31).

The use of whole genome sequencing allowed an unbiased detection of all gene mutations and it is a powerful strategy to uncover genes implicated in the disease. Using this approach we identified a total of 1196 gene mutations in PC346C. Approximately one third of these mutations occurred in nucleotide repeat sequences (see also ref (43)). The total number of gene mutations detected in G089 was 1.5 fold lower than in PC346C, which fits to the mutational phenotype due to MMR deficiency of PC346C. Mutations in MNRs were enriched by 6-fold in PC346C as compared to G089. The bias in MNR mutations in PC346C as compared to G089 was used to select for gene mutations triggered by the MMR system. Several of the downstream target genes of MMR deficiency have shown to be key players in proliferation and apoptosis pathways.

Identification of novel MMR target genes in PCa

A list of 381 genes showing mutations in mononucleotide repeats in PC346C was used to select for novel target gene mutations caused by deficiency in the MMR system in PCa. Since the percentage of PCa cases with MSI was low we extended the test panel to include ovarian, endometrial and colorectal cancer cell lines. A final list of 13 candidate genes was further assessed for their mutation status in cell lines with and without microsatellite instability. *CEP164, PHACTR4* and *MLL3* are genes already known as associated with microsatellite instability and their mutation frequencies were compared to those of the novel gene candidates. We have first selected for the genes with the highest mutation frequencies in our prostate cancer cell line panel. Although *TSHZ2* showed the highest mutation frequencies for colorectal and ovarian cancer cell lines only one out of the four MSI prostate cancer cell lines was positive for the *TSHZ2* mutation. *CNOT1* was mutated in all MSI prostate cancer cell lines but this was, in all

cases, a heterozygous mutation and occurred closed to the end of the coding region. Both *PRRT2* and *DAB2IP* were mutated in three out of the four MSI prostate cancer cell lines. Whereas *PRRT2* mutations were also frequent in both colorectal, endometrial and ovarian cancer cell lines *DAB2IP* mutations were mostly restricted to prostate and colorectal cancer samples. *DAB2IP* has recently been identified as a regulator of tumor growth and apoptosis. The expression of *DAB2IP* is often downregulated in PCa and this downregulation causes activation of the RAS signaling cascade and inactivation of the ASK1-JNK pathway leading to growth stimulation and suppression of apoptosis (44, 45). The observation that *DAB2IP* is mutated at high frequency in MMR deficient prostate and colorectal cell lines indicates a link between MMR and this gene. But so far, epigenetic suppression was suggested as the major mechanism behind *DAB2IP* downregulation (46). Truncating mutations in *PRRT2* were first identified in 2011 by Chen WJ et al. in paroxysmal kinesigenic dyskinesia (PKD) (34). Although *PRRT2* appears to be an important factor in several neurological-related conditions (47) there is no correlation between *PRRT2* and other diseases. Also, there is little functional data regarding the *PRRT2* protein with only one protein-protein interaction known so far with the synaptosomal-associated protein 25 (SNAP25) (48).

PRRT2 and mismatch repair deficiency

The mutation status of *PRRT2* and *DAB2IP* was assessed in a cohort of 24 colorectal and 24 endometrial MSI patient samples. We observed a high number of patients with mutations in *PRRT2* and to a lesser extent mutations in *DAB2IP* were detected. The finding that *DAB2IP* was mutated in only one endometrial cancer sample indicates tissue specificity for this particular gene mutation. *PRRT2* was frequently mutated in both colorectal and endometrial cancer patients. The frequency of *PRRT2* mutation in patient samples is only second to the frequency of *TGFβRII* in colorectal cancer and the highest in endometrial cancer of all MSI genes (Table S7) (49). Although most targets of MMR deficiency are considered to be tumor suppressor genes that through the frameshift outcome lose their function there are also studies showing activating mutations such as the frameshifts in the TCF4 gene (50). The two *PRRT2* frameshift mutations found in PC346C, the insertion of one or two Cs, occur in a C9 repeat in the second coding exon and introduce a premature stop codon in the mRNA. As a consequence, the PC346C cell line lacks a wild type *PRRT2* protein. Although it is claimed that the non-sense mediated mRNA decay (NMR) mechanisms recognize premature stop codons and eliminate aberrant transcripts (51) we found a high level of expression of the *PRRT2* transcript in the PC346C cell line. The expression level of *PRRT2* is overall lower in cancer as compared to normal, which would indicate it as a potential tumor suppressor gene. However mutations in this gene only affect one of the copies in almost all affected cancers, leaving a wild type *PRRT2* copy intact. The PC346C prostate cancer cell line and patient from which the cell line is derived are the only samples with homozygous mutations in the 9C repeat. Despite the lower expression of *PRRT2* in cancer versus normal tissue our data suggests a dominant negative effect of the truncated *PRRT2* protein.

PRRT2 in proliferation and migration

To investigate the functional role of both the wild-type ($PRRT2^{wt}$) and the truncated *PRRT2* protein (ΔPRRT2) we generated stable $PRRT2^{wt}$ and ΔPRRT2 expressing cell lines. Two normal prostate cell lines (PNT2C2 and RWPE-1) and one colorectal cancer cell line (HCT116) were used to assess the effect of *PRRT2* in proliferation, apoptosis and migration. Remarkably, we observed an oncogenic effect of ΔPRRT2. Both proliferation and migration were enhanced by stable ΔPRRT2 expression. Conversely, $PRRT2^{wt}$ appeared to decrease the proliferation rate, being this effect more evident in the PNT2C2 and HCT116 cell lines as compared to RWPE-1. This decrease in proliferation by $PRRT2^{wt}$ overexpression suggests *PRRT2* to participate in cell cycle regulation. The frameshift mutation creates a truncated *PRRT2* protein that increases both proliferation and migration in cell lines with a wild type *PRRT2* protein. Based on our findings, we hypothesize that wild type *PRRT2* inhibits slightly proliferation, whereas truncated *PRRT2* can stimulate, and overrules wild type *PRRT2*. In this way it can function as an oncogene by promoting both proliferation and migration. Although genes are usually dichotomized in either oncogenes or tumor suppressors (52) many can actually exert both functions such as the TP53 tumor suppressor gene which can lead to tissue invasion, metastasis and increased proliferation when mutated (53).

In conclusion, a focused genome-wide sequencing approach, followed by subsequent expression and functional studies indicate mutated *PRRT2* as a novel dominant oncogene, and wt*PRRT2* as a candidate tumor suppressor gene. The finding that the same *PRRT2* mutation detected in this study has also been documented in colorectal, pancreatic and stomach cancer samples (depicted as A214P in Suppl. Fig. 6) supports a more general role of *PRRT2* in cancer biology.

Materials and Methods

DNA samples used in whole genome and data analysis

The DNA of two prostate cancer samples, the PC346C prostate cancer cell line (32) and the G089 prostate cancer patient, were sequenced by Complete Genomics (Complete Genomics Inc, CA). Tissue sample G089 and other patient materials described below has been collected according to national legislation concerning ethical requirements. Use of the clinical samples has been approved by the Erasmus MC Medical Ethics Committee according to the Medical Research Involving Human Subjects Act (MEC-2004-261). NCBI build 36 (hg18) was used as a reference genome during the mapping and data analysis process. Complete Genomics Inc. pipeline generated a report with all single nucleotide variants present in the two samples. Further information on the Complete Genomics Inc. data processing is present in supplementary materials and methods.

DNA from additional cell lines and tissue samples

Additional genomic DNA was isolated from MMR-proficient prostate cancer cell lines PC3, PC135 and PC295 and MMR-deficient prostate cancer cell lines DU145, LNCaP and PC374. Cell pellets from thirteen colorectal, seven endometrial and three ovarian cancer cell lines were kindly provided by Dr. W. Dinjens, Dept Pathology, Erasmus MC, Rotterdam, The Netherlands. Also, genomic DNA was isolated from frozen 31 primary prostate cancer patient tissues, 37 transurethral resection of the prostate (TURP) samples and 11 lymph node prostate cancer metastasis samples. Genomic DNA from paraffin embedded formalin fixed tissue samples of 24 microsatellite instable colorectal cancer patients and 24 microsatellite instable endometrial cancer patients were kindly provided by Dr. W. Dinjens.

DNA isolation

DNA isolation from prostate cancer samples was performed using the QIAamp DNA Blood Midi Kit (Qiagen) according to the manufacturers' instructions. Cell pellets were resuspended in 1 ml PBS (BioWhittaker) and instructions were followed according to the protocol used for 1 ml whole blood. DNA was eluted in 200 µl elution buffer. DNA isolation from the colorectal, endometrial and ovarian cancer cell lines pellets was performed using the Gentra Puregene Cell Kit (Qiagen) (approx. 15 µg DNA from 2x106 cells) following the manufacturer's protocol. Concentration and purity of the DNA was assessed using the NanoDrop ND 1000 spectrophotometer (Nanodrop) by absorption measurements at 260 nm. DNA was stored at -20 degrees Celsius.

RNA isolation

Total RNA was isolated from PC346C, RWPE-1, VCaP, HCT116 and PNT2C2 using the RNeasy kit (Qiagen) according to the manufacturer's protocol. RNA was eluted in 50 µl RNase free water. Concentration and purity of RNA was assessed using the NanoDrop ND 1000 spectrophotometer (Nanodrop) by absorption measurements at 260 nm. RNA was stored at -80 degrees Celsius.

Cell culture

The PC346C cell line was cultured in DMEM-F12 (BioWhittaker), supplemented with 2% (V/V) FCS (PAN Biotech), 1% insulintransferrin-selenium (GIBCO BRL), 0.01% BSA (Boehringer-Mannheim), 10 ng/ml epidermal growth factor (Sigma-Aldrich) and 500 U penicillin-streptomycin (BioWhittaker), 100 ng/ml bronectin (Harbor Bio Products), 20 mg/ml fetuine (ICN Biomedicals), 0.1 nM R1881 (Sigma-Aldrich), 50 ng/ml cholera toxin (Sigma-Aldrich), 0.1 mM phosphoethanolamine (Sigma-Aldrich), 0.6 ng/ml triodothyronine (Sigma-Aldrich), and 500 ng/ml dexamethasone (Sigma-Aldrich). The PC346C cell line expresses the wild-type androgen receptor and secretes high levels of PSA. RWPE-1 cells were cultured in keratinocyte medium (GIBCO BRL), supplemented with 5 ng/ml epidermal growth factor, 1% penicillin-streptomycin (BioWhittaker), and 50 mg/L bovine pituitary

extract. RWPE-1 is a normal prostate epithelium cell line that is androgen-independent and expresses both the androgen receptor and PSA. The colorectal cancer cell line HCT116 and the normal prostate cell line PNT2C2 cells were cultured in RPMI-1640 (GIBCO BRL), supplemented with 10% FCS (PAN Biotech) and 1% penicillin-streptomycin (BioWhittaker). All cell lines were cultured at 37 degrees in a 5% CO_2 atmosphere.

Mononucleotide repeat analysis

A microsatellite instability assay was used to detect changes in the length of mononucleotide repeats by PCR amplification. Primers were designed using the Primer3Plus program. The final fragment size was set to vary between 100 and 130 bp in order to allow precise fragment sizing and robust amplification of all the genes tested. Primer sequences are described in Supplementary materials and methods. The M13 sequence (GTAAAACGACGGCCAGT) was added at the 5' end of each forward primer. This enabled the use of a M13-FAM primer to fluorescently label each PCR fragment. PCR reactions were performed in a total volume of 16 µl containing 12.5 ng genomic DNA, 3 µl 10x PCR Buffer (Qiagen), 2 µl 5x Q-Solution (Qiagen), 0.9 µl 25 mM MgCl2 (Quiagen), 0.3 µl dNTPs (10 mM), 0.5 µl forward primer (100 ng/µl), 0.5 µl M13-FAM primer (100 ng/µl), 1 µl reverse primer (100 ng/µl), and 1 U HotStarTaq Polymerase (Qiagen). An initial denaturation step of 15 min at 95 degrees Celsius was used to activate the HotstarTaq, followed by 35 cycles denaturation at 95 degrees Celsius for 30 sec, an annealing step at 54 degrees Celsius for 30 sec and elongation at 72 degrees Celsius for 1 min. A final elongation step at 72 degrees Celsius for 10 min was used. PCR products were checked by electrophoresis in a 1% agarose gel.

Fragment analysis

The PCR products were separated by capillary electrophoresis using an ABI PRISM 3130 sequencing analyzer. This method allows single base pair resolution of PCR fragments. A mixture of 1.5 µl of the PCR products with 9.8 µl highly-deionized formamide (HiDi) and 0.2 µl GeneScan-500 LIZ size standard (Applied Biosystems) was prepared. This mixture was heated for 2 min at 95 degrees Celsius and subsequently loaded to the sequence analyzer. Data was analyzed with the Peak Scanner software (Applied Biosystems).

Sequencing

PCR products have been sequenced bi-directionally using standard Sanger sequencing. The sequencing reactions were carried out using the same forward and reverse primers as indicated above in the PCR but at different concentrations (3 ng primer per reaction). The final reaction volume was 20 µl. The sequencing products were precipitated using isopropanol and resuspended in 20 ul formamide (Sigma-Aldrich). The PCR product was sequenced on an ABI Model 3730 automated sequencer and analyzed using DNAMAN (Lynnon Corporation).

PRRT2 expression constructs

The sequence validated I.M.A.G.E clone IRATp970E0579D corresponding to the BC053594 clone accession number was purchased from Source Bioscience, Nottingham, UK. This plasmid contains the wt*PRRT2* cDNA sequence. Primers 5' GAT CGA ATT CGT TTG CCG CTG TCT CT 3' (Fw) and 5' GAT CGC GGC CGC TCA CTT ATA CAC GCC 3' (Rv) were used to amplify the wt*PRRT2* cDNA and subsequently clone the fragment into the pCDH-CMV-MCS-EF1-GFP-T2A-Puro vector (System Biosciences). Primers 5' GAT CGA ATT CGT TTG CCG CTG TCT CT 3' (Fw) and 5' GAT CGC GGC CGC CCC TTC TCA TTC GAT 3' (Rv) were used to generate a truncated *PRRT2* cDNA fragment. This fragment was integrated into the pCDH-CMV-MCS-EF1-GFP-T2A-Puro vector (System Biosciences) for expression of ΔPRRT2 protein. The amplified fragments were ligated to the vector using the Rapid DNA Ligation Kit according to the manufacturer's protocol (Roche). JM109 competent bacteria (Promega) were transformed with the ligation products using the heat shock method and plated in Ampicillin LB agar plates. DNA isolation was performed on selected colonies (Quiagen) and restriction enzyme reactions (BamH1 and Not1) confirmed the proper ligation of fragment to vector.

Lentivirus production

HEK293T cells were co-transfected with pCDH-wtPRRT2-GFP-Puro or pCDH-ΔPRRT2-GFP-Puro or pCDH-GFP-Puro expression vectors and pPAX2 and pMD2.G (kind gift of Prof Didier Trono, Switserland) using the calcium phosphate precipitation method. Virus was harvested from the supernatant and used for transduction of PNT2C2, RWPE-1 and HCT116 cells. Pools of infected cells were propagated.

Functional role of PRRT2

The functional role of PRRT2 was assessed in RWPE-1, HCT116 and PNT2C2 cell lines expressing the truncated form of *PRRT2* (ΔPRRT2), the wildtype *PRRT2* (wt*PRRT2*) and the control GFP. These cell lines were propagated in medium supplemented with 1 µg/mL puromycin (Millipore). Briefly, cells were plated in 96-well culture plates at a density of 1000 cells per well (100 µl). Proliferation was assessed using the CellTiter-Glo Luminescent Cell Viability Assay (Promega), according to manufacturer's protocol. Apoptosis was assessed using the ApoLive-Glo™ Multiplex Assay (Promega), which measures both the number of viable cells as a marker of cytotoxicity and caspase activation as a marker of apoptosis within a single assay well. The assay was performed according to the manufacturer's protocol. As positive control of apoptosis an apoptosis inducer set (Millipore) was used. A mix containing 700x dilutions of Actinomycin D (10 mM), Camptothecin (2 mM), Cycloheximide (100 mM), Dexamethasone (10 mM) and Etoposide (10 mM) was supplemented for 24h to induce apoptosis. The QCM Chemotaxis Cell Migration Assay, 96-well (8 µm), fluorimetric (Millipore) was performed according to the manufacturer's protocol.

Acknowledgements

We would like to thank Wytske van Weerden, Erasmus MC for PC346C cell line model systems, Complete Genomics Inc. for assistance with the NGS data, Arno van Leenders, Erasmus MC for patient sample selection. This research was made possible by financial contributions from CTMM, project PCMM (project number 03O-203), the FP7 Marie Curie Initial Training Network PRO-NEST (grant number 238278) and the Foundation for Scientific Urological Research (SUWO).

References

1. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008;18(1):85-98.
2. Shah SN, Hile SE, Eckert KA. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer research. 2010;70(2):431-5.
3. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. Trends in biotechnology. 2007;25(11):490-8.
4. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature reviews Genetics. 2004;5(6):435-45.
5. Pena-Diaz J, Jiricny J. Mammalian mismatch repair: error-free or error-prone? Trends in biochemical sciences. 2012;37(5):206-14.
6. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453(7192):175-83.
7. Loeb LA. Human cancers express mutator phenotypes: origin, consequences and targeting. Nature reviews Cancer. 2011;11(6):450-7.
8. Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science. 1995;268(5215):1336-8.
9. Souza RF, Appel R, Yin J, Wang S, Smolinski KN, Abraham JM, et al. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. Nature genetics. 1996;14(3):255-7.
10. Guanti G, Resta N, Simone C, Cariola F, Demma I, Fiorente P, et al. Involvement of PTEN mutations in the genetic pathways of colorectal cancerogenesis. Human molecular genetics. 2000;9(2):283-7.
11. Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, Reed JC, et al. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science. 1997;275(5302):967-9.
12. Schwartz S, Yamamoto H, Navarro M, Maestro M, Reventós J, Perucho M. Frameshift Mutations at Mononucleotide Repeats in caspase-5 and Other Target Genes in Endometrial and Gastrointestinal Cancer of the Microsatellite Mutator Phenotype. Cancer research. 1999;59(12):2995-3002.
13. Riccio A, Aaltonen LA, Godwin AK, Loukola A, Percesepe A, Salovaara R, et al. The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. Nature genetics. 1999;23(3):266-8.
14. Duval A, Gayet J, Zhou XP, Iacopetta B, Thomas G, Hamelin R. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. Cancer research. 1999;59(17):4213-5.
15. Alazzouzi H, Davalos V, Kokko A, Domingo E, Woerner SM, Wilson AJ, et al. Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. Cancer research. 2005;65(22):10170-3.
16. Liu WG, Dong XY, Mai M, Seelan RS, Taniguchi K, Krishnadath KK, et al. Mutations in AXIN2 cause colorectal cancer with defective mismatch repair by activating beta-catenin/TCF signalling (vol 26, pg 146, 2000). Nature genetics. 2000;26(4):501-.

17. Hempen PM, Zhang L, Bansal RK, Iacobuzio-Donahue CA, Murphy KM, Maitra A, et al. Evidence of selection for clones having genetic inactivation of the activin A type II receptor (ACVR2) gene in gastrointestinal cancers. Cancer research. 2003;63(5):994-9.
18. Ionov Y, Yamamoto H, Krajewski S, Reed JC, Perucho M. Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. P Natl Acad Sci USA. 2000;97(20):10872-7.
19. Cuilliere-Dartigues P, El-Bchiri J, Krimi A, Buhard O, Fontanges P, Flejou JF, et al. TCF-4 isoforms absent in TCF-4 mutated MSI-H colorectal cancer cells colocalize with nuclear CtBP and repress TCF-4-mediated transcription. Oncogene. 2006;25(32):4441-8.
20. de la Chapelle A. Genetic predisposition to colorectal cancer. Nature Reviews Cancer. 2004;4(10):769-80.
21. Umar A, Risinger JI, Hawk ET, Barrett JC. Guidelines - Testing guidelines for hereditary non-polyposis colorectal cancer. Nature Reviews Cancer. 2004;4(2):153-8.
22. Pal T, Permuth-Wey J, Kumar A, Sellers TA. Systematic Review and Meta-analysis of Ovarian Cancers: Estimation of Microsatellite-High Frequency and Characterization of Mismatch Repair Deficient Tumor Histology. Clin Cancer Res. 2008;14(21):6847-54.
23. Nelson GS, Pink A, Lee S, Han GM, Morris D, Ogilvie T, et al. MMR deficiency is common in high-grade endometrioid carcinomas and is associated with an unfavorable outcome. Gynecol Oncol. 2013;131(2):309-14.
24. Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, Sougnez C, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497(7447):67-73.
25. Leite M, Corso G, Sousa S, Milanezi F, Afonso LP, Henrique R, et al. MSI phenotype and MMR alterations in familial and sporadic gastric cancer. Int J Cancer. 2011;128(7):1606-13.
26. Imai K, Yamamoto H. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. Carcinogenesis. 2008;29(4):673-80.
27. Leach FS, Velasco A, Hsieh JT, Sagalowsky AI, McConnell JD. The mismatch repair gene hMSH2 is mutated in the prostate cancer cell line LNCaP. J Urology. 2000;164(5):1830-3.
28. Chen Y, Wang JS, Fraig MM, Metcalf J, Turner WR, Bissada NK, et al. Defects of DNA mismatch repair in human prostate cancer. Cancer research. 2001;61(10):4112-21.
29. Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. P Natl Acad Sci USA. 2011;108(41):17087-92.
30. Pritchard CC, Morrissey C, Kumar A, Zhang XT, Smith C, Coleman I, et al. Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. Nat Commun. 2014;5.
31. Alves IT, Hartjes T, McClellan E, Hiltemann S, Bottcher R, Dits N, et al. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. Oncogene. 2015;34(5):568-77.
32. Marques RB, Erkens-Schulze S, de Ridder CM, Hermans KG, Waltering K, Visakorpi T, et al. Androgen receptor modifications in prostate cancer cells upon long-termandrogen ablation and antiandrogen treatment. International Journal of Cancer. 2005;117(2):221-9.
33. Woerner SM, Yuan YP, Benner A, Korff S, von Knebel Doeberitz M, Bork P. SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. Nucleic acids research. 2010;38(Database issue):D682-9.

34. Chen W-J, Lin Y, Xiong Z-Q, Wei W, Ni W, Tan G-H, et al. Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. Nature genetics. 2011;43(12):1252-5.
35. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.
36. Wierdl M, Greene CN, Datta A, Jinks-Robertson S, Petes TD. Destabilization of simple repetitive DNA sequences by transcription in yeast. Genetics. 1996;143(2):713-21.
37. Peltomaki P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Human molecular genetics. 2001;10(7):735-40.
38. Kagan J, Pisters L, Troncoso P, Joe Y, Parat J, Babaian R, et al. Genetic instability in microsatellite sequences in prostate-cancer. International journal of oncology. 1994;5(4):921-4.
39. Watanabe M, Imai H, Shiraishi T, Shimazaki J, Kotake T, Yatani R. Microsatellite instability in human prostate cancer. British journal of cancer. 1995;72(3):562-4.
40. Azzouzi AR, Catto JW, Rehman I, Larre S, Roupret M, Feeley KM, et al. Clinically localised prostate cancer is microsatellite stable. BJU international. 2007;99(5):1031-5.
41. Dahiya R, Lee C, McCarville J, Hu W, Kaur G, Deng G. High frequency of genetic instability of microsatellites in human prostatic adenocarcinoma. International journal of cancer Journal international du cancer. 1997;72(5):762-7.
42. Perinchery G, Nojima D, Goharderakhshan R, Tanaka Y, Alonzo J, Dahiya R. Microsatellite instability of dinucleotide tandem repeat sequences is higher than trinucleotide, tetranucleotide and pentanucleotide repeat sequences in prostate cancer. International journal of oncology. 2000;16(6):1203-9.
43. Kim T-M, Laird Peter W, Park Peter J. The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. Cell. 2013;155(4):858-68.
44. Xie D, Gore C, Zhou J, Pong RC, Zhang H, Yu L, et al. DAB2IP coordinates both PI3K-Akt and ASK1 pathways for cell survival and apoptosis. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(47):19878-83.
45. Min J, Zaslavsky A, Fedele G, McLaughlin SK, Reczek EE, De Raedt T, et al. An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-kappaB. Nature medicine. 2010;16(3):286-94.
46. Dote H, Toyooka S, Tsukuda K, Yano M, Ota T, Murakami M, et al. Aberrant promoter methylation in human DAB2 interactive protein (hDAB2IP) gene in gastrointestinal tumour. British journal of cancer. 2005;92(6):1117-25.
47. Meneret A, Gaudebout C, Riant F, Vidailhet M, Depienne C, Roze E. PRRT2 mutations and paroxysmal disorders. European journal of neurology : the official journal of the European Federation of Neurological Societies. 2013;20(6):872-8.
48. Lee HY, Huang Y, Bruneau N, Roll P, Roberson ED, Hermann M, et al. Mutations in the gene PRRT2 cause paroxysmal kinesigenic dyskinesia with infantile convulsions. Cell reports. 2012;1(1):2-12.
49. Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer research. 2002;62(9):2447-54.
50. Duval A, Rolland S, Tubacher E, Bui H, Thomas G, Hamelin R. The human T-cell transcription factor-4 gene: structure, extensive characterization of alternative splicings, and mutational analysis in colorectal cancer cell lines. Cancer research. 2000;60(14):3872-9.

51. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nature structural & molecular biology. 2009;16(2):107-13.
52. Lou X, Zhang J, Liu S, Xu N, Liao DJ. The other side of the coin: The tumor-suppressive aspect of oncogenes and the oncogenic aspect of tumor-suppressive genes, such as those along the CCND–CDK4/6–RB axis. Cell cycle. 2014;13(11):0--1.
53. Muller PA, Vousden KH. p53 mutations in cancer. Nature cell biology. 2013;15(1):2-8.

Supplementary data

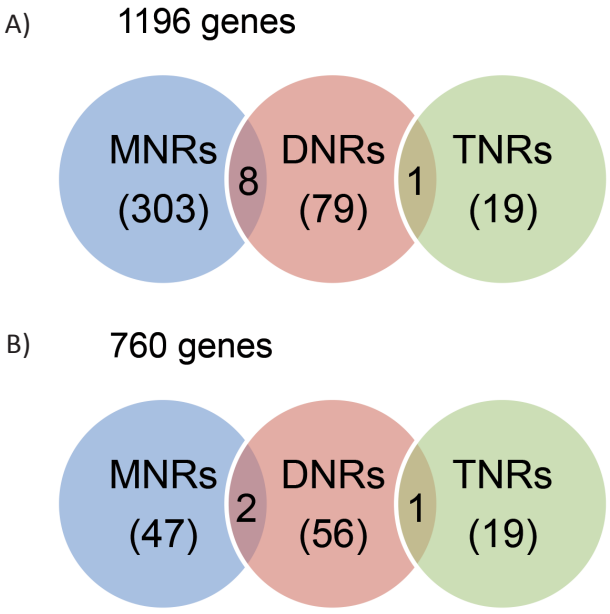


Figure S1: Distribution of mononucleotide (MNR), dinucleotide (DNR) and trinucleotide (TNR) repeat mutations of length 6 nucleotides of MNRs and DNRs and length of 9 nucleotides for TNRs in affected genes. (A) represents the distribution in the PC346C DNA and (B) in G089 DNA. The numbers in the intersection of the different categories represent genes with both NR (either MNR and DNR or DNR and TNR). No genes had both MNR and TNR.

Table S3: Distribution of mutated genes in PC346C and G089 and in all UCSC hg18 genes in MNR categories of different type and length. The subscript number represents the total nucleotide length of the repeat.

	PC346C (1196 genes)						G089 (760 genes)						UCSC genes					
	n_6	n_7	n_8	n_9	n_{10}	n_{12}	n_6	n_7	n_8	n_9	n_{10}	n_{12}	n_6	n_7	n_8	n_9	n_{10}	n_{12}
A	96	78	54	29	19	4	13	11	7	4	4	0	15100	9501	5768	4025	3013	1880
T	84	66	44	25	15	1	12	7	5	2	2	0	14617	9690	6358	4672	3616	2359
C	68	33	15	5	1	0	10	6	4	0	0	0	8636	2106	522	225	128	44
G	55	26	8	3	1	0	12	9	5	3	2	0	7468	1630	423	157	83	24

Table S4: Allele pattern of microsatellite instability gene targets in MMR-deficient and -proficient prostate cancer cell lines. Genes are ordered first by the frequency of mutation in MMR-deficient prostate cancer samples and secondly alphabetically. The three MMR associated genes used as positive controls are depicted in red. MSS stands for microsatellite stable and MSI for microsatellite instable. WT represents the wildtype allele. M1 and M2 correspond to deletion of one or two nucleotides, respectively. P1 and P2 correspond to insertion of one or two nucleotides.

	MMR-proficient (MSS)				MMR-deficient (MSI)		
Gene	**PC3**	**PC295**	**PC135**	**DU145**	**LNCaP**	**PC346C**	**PC374**
CNOT1	wt	wt	wt	m1wt	m1m2wt	m1m2wt	m1wt
DAB2IP	wt	wt	wt	wt	m1wt	m1wt	wtp1
PRRT2	wt	wt	wt	wtp1p2	wt	p1p2	wtp1
TTC3	wt	wt	wt	m1wt	wt	m1wt	m1wt
CEP164	wt	wt	wt	wt	m1wt	m1wt	m1wt
PHACTR4	wt	wt	wt	wt	m1wt	m1wt	m1wtp1
TROVE2	wt	wt	wt	wt	wt	m1wt	m1wt
ZFR	wt	wt	wt	wt	m1wt	m1wt	wt
MLL3	wt	wt	wt	wt	m1wt	m1wt	wt
ANLN	wt	wt	wt	wt	wt	m1wt	wt
ANUBL1	wt	wt	wt	wt	wt	m1wt	wt
EPRS	wt	wt	wt	wt	wt	wtp1	wt
KCNMA1	wt	wt	wt	wt	wt	m1wt	wt
PDS5A	wt	wt	wt	wt	wt	m1wt	wt
SFRS12IP1	wt	wt	wt	wt	wt	wtp1	wt
TSHZ2	wt	wt	wt	wt	wt	m1wt	wt
USP42	wt	wt	wt	wt	wt	m1wt	wt

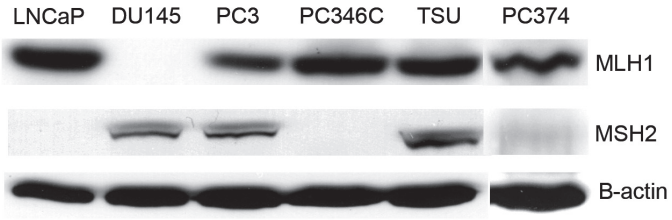


Figure S2: Western blot analysis of the MLH1 and MSH2 protein expression in the MMR deficient prostate cancer cell lines LNCaP, DU145 and xenograft transplant PC374 and in the MMR proficient prostate cell lines PC3 and TSU. Actin is displayed as a loading control.

Table S5: Mutation profiling of the selected gene panel in all colorectal, endometrial and ovarian cancer cell lines used to assess the mutation status. The MMR proficient cell lines are depicted in blue and the MMR deficient cells lines are in orange. ND represents not defined. WT represents the wildtype. M1, M2 and M3 correspond to deletion of one, two or three nucleotides, respectively. P1 and P2 correspond to insertion of one or two nucleotides.

	Colorectal Cancer											
Gene	**CACO-2**	**HT-29**	**SW-480**	**HCT116**	**LOVO**	**LS174T**	**CO115**	**RKO**	**TC7**	**TC71**	**4224**	**5583S**	**DV-90**
PRRT2	wt	wt	wt	wt	m1wtp1p2	m1wt	m1wt	wt	wt	wt	wt	wt	m1wt
TSHZ2	wt	wt	wt	m1wt	m1wt	m1wt	m1wt	m1wt	m1wt	m1wt	m1wt	m1wt	m1wt
ANLN	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
CEP164	wt	wt	wt	m1	m1wt	m1wtp1	m1	m1	m1wt	m1wt	wt	wt	m1wt
CNOT1	wt	wt	ND	m1m2	m1m2	m1m2	m1m2m3	m1m2	m1m2	m1m2	wt	m1m2	m1m2
KCNMA1	wt	wt	wt	m1m2wt	m1m2wt	m1m2wt	m1m2wt	m1m2wt	wt	wt	wt	wt	m1m2wt
ANUBL1	wt	wt	wt	wt	wt	wt	wt	wt	wt	m1wt	wt	m1wt	wt
EPRS	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
TTC3	wt	wt	wt	m1wt	m1wt	wt	m1wt	m1wt	wt	m1wt	wt	ND	ND
SFRS12IP1	wt	wt	wt	wt	wt	m1	m2m1wt	wt	m1	m1	wt	m1wt	m1wt
ZFR	wt	wt	wt	wt	wt	wt	wt	m1wt	m1wt	m1wt	wt	wt	m1wt
DAB2IP	wt	wt	wt	m1wt	m1wt	wt	m1p1	m1wt	m1p1	wtp1	wt	wt	wt
PHACTR4	wt	wt	wt	m1wt	m1wt	m1wt	m2	m1wt	ND	wtp1	wt	m1wt	wt
MLL3	wt	wt	wt	m1wt	m1wt	wt	m1wt	m1wt	m1wt	m1wt	wt	m1wt	m1wt
PDS5A	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
TROVE2	wt	wt	wt	m1wt	wt	wt	m1wt	wt	m1wt	wt	wt	wt	m1wt
USP42	wt	wt	wt	wt	wt	wt	m1wt	wt	m1wt	m1wt	wt	wt	m1wt

Table S5: Continued

	Endometrial Cancer							Ovarian Cancer				
Gene	**ECC-1**	**SCRC**	**KLE**	**HEC59**	**AN3CA**	**HEC1B**	**RL95-2**	**CAOV-3**	**SKOV-6**	**SKOV3**	**OVCAR3**	**2774**
PRRT2	wt	wt	wt	m1wt	wtp1p2	m1wt	m1wt	wt	wt	wt	wt	m1wtp1
TSHZ2	wt	wt	wt	m1wt	m1wt	m1wt	ND	wt	wt	m1wt	m1wt	m1wt
ANLN	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
CEP164	wt	wt	wt	wt	m1wt	m1wt	m1	wt	wt	m1wt	wt	m1
CNOT1	wt	wt	wt	m1m2	m1wt	m1wt	m1m2m3	wt	wt	wt	wt	m1wt
KCNMA1	wt	wt	wt	wt	wt	wt	m1m2wt	wt	wt	wt	wt	m1m2wt
ANUBL1	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
EPRS	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt	wt
TTC3	ND	ND	wt	ND	m1wt	wt	wt	wt	wt	wt	wt	ND
SFRS12IP1	wt	wt	wt	m1wt	wt	wt	wtp1	wt	wt	wt	wt	m1
ZFR	wt	wt	wt	m1wt	m1wt	m1wt	wt	wt	wt	wt	wt	m1wt
DAB2IP	wt	wt	wt	m1wt	wt	wt	wt	wt	wt	wt	wt	wt
PHACTR4	wt	wt	wt	m1wt	m1wt	wt	m1wt	wt	wt	wt	wt	m1wt
MLL3	wt	wt	wt	wt	m1wt	m1wt	m1wt	wt	wt	wt	wt	m1m2
PDS5A	wt	wt	wt	wt	m1wt	wt	wt	wt	wt	wt	wt	wt
TROVE2	wt	wt	wt	m1	m1wt	m1wt	m1wt	wt	wt	m1wt	wt	wt
USP42	wt	wt	wt	m1wt	m1wt	wt	m1wt	wt	wt	wt	wt	wt

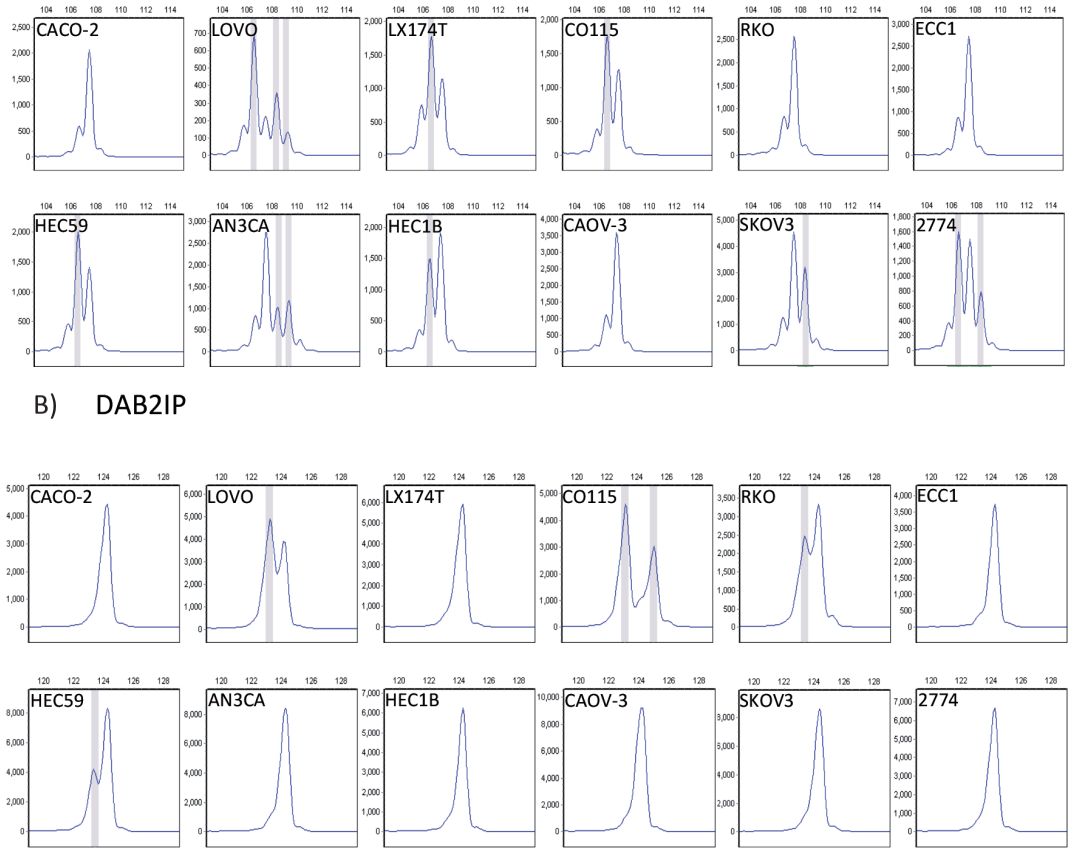


Figure S3: Analysis of *PRRT2* (A) and *DAB2IP* (B) PCR fragments containing repeat sequences (see also Figure 2) in colorectal, endometrial and ovarian cancer cell lines. CACO2 is MMR proficient colon cancer cell line; LOVO, LX174T, CO115 and RKO are MMR deficient colon cancer cell lines, respectively. ECC1 is a MMR proficient endometrial cancer cell line; HEC59, AN3CA and HEC1B are MMR deficient endometrial cancer cell lines, respectively. CAOV-3 is a MMR proficient ovarian cancer cell line; SKOV3 and 2774 are MMR deficient ovarian cancer cell lines, respectively. The highlighted peak represents a mutant allele. The Y-axis corresponds to the intensity of the fluorescent peak. The X-axis represents the fragment size in base pairs. Shoulder peaks with intensities less than 32% of the highest peak were discarded from further analysis. PCR primers are given in Supplementary Materials and Methods.

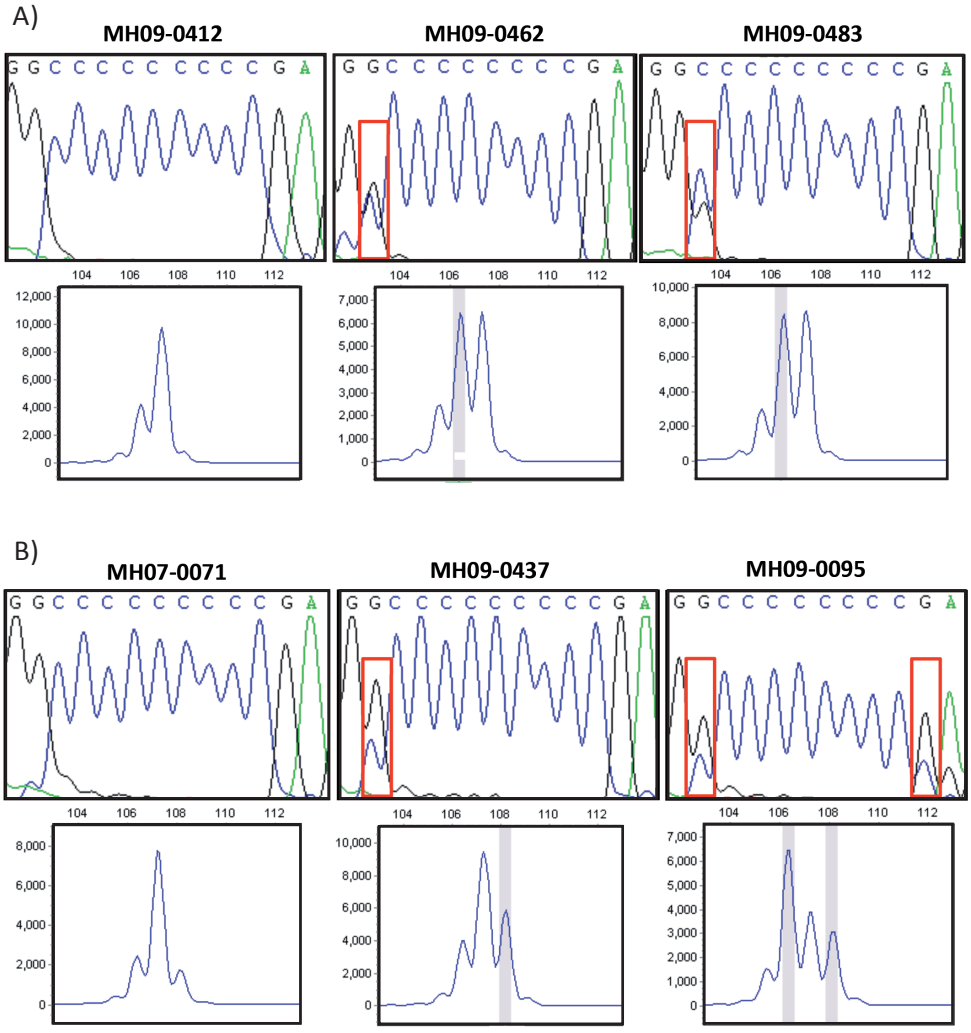


Figure S4) Analysis of the *PRRT2* PCR fragment containing the C9 repeat sequence and repeat sequence analysis of MSI colorectal and endometrial cancer patient samples. (A) Chromatogram and fragment analysis of three colorectal cancer patients. (B) Chromatogram and fragment analysis of three endometrial cancer patients. The highlighted peaks (grey outline and red circle) represent mutant alleles. The wildtype 9 C repeat shows a prominent 107 bp fragment and a shoulder peak at length 106 bp. The intensity of shoulder peaks are less than 32% the intensity of the highest peak. Both MH09-0462 and MH09-0483 display peaks of equal intensity at length 106 and 107. The peak at 106 bp is therefore highlighted as a mutant allele in the two samples mentioned above. In (B) MH09-0095 has a lower peak corresponding to the wt 107 fragment. Fragments of 106 and 108 bp are also present in this sample.

Table S6) Allele pattern of *PRRT2* and *DAB2IP* in MSI colorectal and endometrial cancer patients. ND represents not defined. WT represents the wildtype allele. M1 and M2 correspond to deletion of one or two nucleotides, respectively. P1 corresponds to insertion of one nucleotide.

Patient	Type of cancer	PRRT2	DAB2IP
MH09-0412	Colon	wt	wt
MH09-0419	Colon	m1wt	wt
MH09-0421	Colon	wt	wt
MH09-0443	Colon	m1	wt
MH09-0456	Colon	wt	wt
MH09-0462	Colon	m1wt	wt
MH09-0467	Colon	wt	m1wt
MH09-0477	Colon	wt	wtp1
MH09-0483	Colon	m1wt	wt
MH09-0496	Colon	wt	wt
MH09-0503	Colon	wt	wt
MH09-0515	Colon	wtp1	wt
MH09-0541	Colon	wt	wt
MH09-0543	Colon	wt	wt
MH09-0569	Colon	wtp1	m1wt
MH09-0575	Colon	wtp1	wt
MH09-0586	Colon	m1wt	wt
MH09-0587	Colon	m1wt	wt
MH09-0609	Colon	m1wt	wt
MH09-0648	Colon	m1wt	wtp1
MH10-0016	Colon	m1wt	wt
MH10-0017	Colon	m1wt	wtp1
MH10-0030	Colon	m1wt	wt
MH10-0122	Colon	m2m1wt	m1wt
MH07-0071	Uterus	wtp1	wt
MH07-0120	Uterus	m1wt	wt
MH08-0225	Uterus	wt	wt
MH08-0249	Uterus	wt	wt
MH08-0265	Uterus	wt	wt
MH08-0318	Uterus	wt	wt
MH09-0060	Uterus	wtp1	wt
MH09-0071	Uterus	ND	ND
MH09-0095	Uterus	wtp1	wt
MH09-0099	Uterus	wt	wt
MH09-0109	Uterus	wt	wt
MH09-0203	Uterus	wt	ND
MH09-0204	Uterus	wt	ND

Table S6) Continued

MH09-0308	Uterus	m1wt	wt
MH09-0363	Uterus	wt	-
MH09-0393	Uterus	wtp1	wt
MH09-0437	Uterus	m1wtp1	wt
MH09-0440	Uterus	wt	m1wt
MH09-0479	Uterus	m1wt	wtp1
MH09-0486	Uterus	wt	wt
MH09-0535	Uterus	m1wt	wt
MH09-0572	Uterus	m1wt	wt
MH09-0656	Uterus	wt	wt
MH09-0657	Uterus	m1wt	-

Table S7) Frequency of mutations in MNRs in known MSI sensitive target genes, and in *PRRT2* and *DAB2IP*. CRC is colorectal cancer, EC is endometrial cancer and CRCc is colorectal cancer cell. A – means not determined.

Gene	CRC %	EC %	CRCc %
TGFβR2	74.6	14.6	88.6
BAX	41.7	24.0	56.5
Caspase 5	47.6	11.4	84.6
PTEN	18.7	11.6	0.0
MSH6	24.3	15.8	37.0
IGFR2	20.0	16.0	22.7
CEP164	38.1	-	-
PHACTR4	23.8	-	70.0
MLL3	47.6	-	70.0
PRRT2	62.5	45.8	40.0
DAB2IP	25.0	8.3	60.0

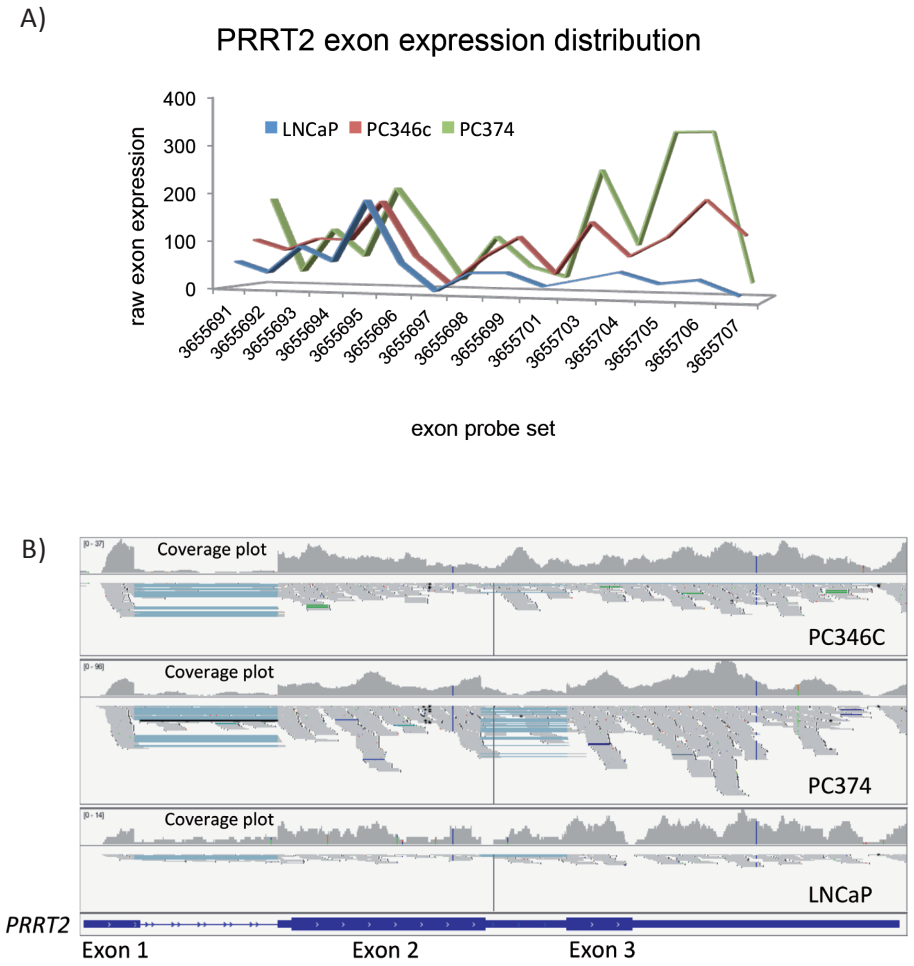
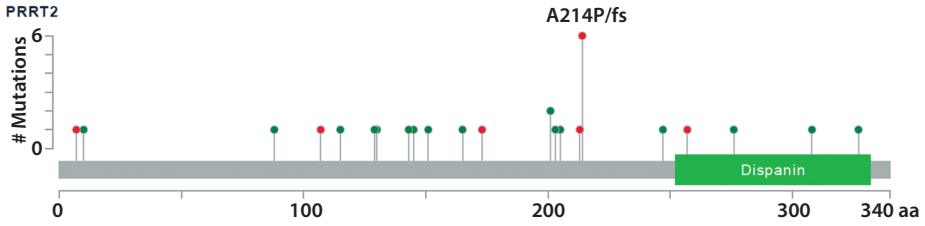


Figure S5: Expression of the *PRRT2* transcript in PC346C, PC374 and LNCaP. (A) Raw expression values of the exon probes across the *PRRT2* transcript. (B) RNAseq read distribution of the *PRRT2* transcript. The reads mapping to *PRRT2* are plotted bellow the coverage plot. The upper panel represents PC346C, the middle panel PC374 and the lower panel LNCaP. *PRRT2* exons are depicted in blue at the end.



Sample ID	Cancer Study	AA change	Type	Copy #	COSMIC	#Mut in Sample
TCGA-AA-3715-01	Colorectal (TCGA pub)	*A214fs*	FS del	diploid	2	1490
TCGA-AA-3715-01	Colorectal (TCGA)	*A214fs*	FS del	diploid	2	1480
MD-339	MBL (Broad)	*A214P*	Missense	NA	2	9
TCGA-HV-A5A5-01	Pancreas (TCGA)	*A214fs*	FS ins	diploid	2	166
TCGA-FZ-5924-01	Pancreas (TCGA)	*A214fs*	FS del	diploid	2	54
TCGA-IB-7654-01	Pancreas (TCGA)	*A214fs*	FS del	gain	2	198
TCGA-BR-7703-01	Stomach (TCGA)	*A214P*	Missense	diploid	2	602

Figure S6: Spectrum of somatic alterations in *PRRT2*. The cBIOPortal for cancer genomics [http://www.cbioportal.org/public-portal/] was used to catalogue all the genetic alterations found in the *PRRT2* gene across several cancer datasets. (A) Diagram of all mutations catalogued in the *PRRT2* gene. Red corresponds to truncating mutations and green to missense mutations. The A214P frameshift corresponds to the mutation also found in the PC346C cell line. (B) List of the samples containing the A214P mutation. This mutation has additionally been detected in 3 pancreatic adenocarcinoma samples, 1 colorectal adenocarcinoma sample, 1 stomach adenocarcinoma sample and 1 medulloblastoma sample.

Supplementary materials and methods

DNA sequencing

Paired-end sequencing of the G089 and PC346C DNA samples was performed with the Complete Genomics service provider using a proprietary sequencing-by-ligation technology and primary data analysis, including image analysis, base calling, alignment and variant calling (1). Reads were mapped to the NCBI Build 36.1 reference genome and mappings were expanded by local de novo assembly on all regions of the genome that contain single nucleotide variations (SNVs) relative to the reference genome (1). SNVs, insertions and deletions (indels) and substitutions are reported in the variation files. Additional information on the nomenclature used by Complete Genomics can be found in 'http://www.completegenomics.com/FAQs/Variant-Calls-SNPs-and-Small-Indels/'.

Public exon array datasets

We used a publicly available dataset of Affymetrix Human Exon Arrays to determine the expression levels of PRRT2. The prostate cancer dataset contains 48 previously published prostate cancer samples (GSE41408, (2)) as well as additional cancerous and control samples, accessible via GEO accession number GSE59745. The datasets comprised samples from normal adjacent prostate (NAP), localized prostate cancer obtained via radical prostatectomy (PCa) and transurethral resection of the prostate (TURP), as well as metastasis in lymph node (LN PCa). Public datasets of lung cancer (GSE12236, (3)) and gastric cancer (GSE13195) were used to confirm the expression pattern of PRRT2.

Table 1: Primer sequences for microsatellite analysis.

Gene symbol	Primer	Sequence: 5' - 3'	Fragment length (bp)
PRRT2	Forward	CTCACTCACCACCCTCAAAAA	107
	Reverse	TTCTCATTCGATCCTCCTCAAC	
TSHZ2	Forward	AAGCACGCTCTGTCTGACATC	130
	Reverse	CTGACATCCATTTCCAGCTTC	
ANLN	Forward	GTGATTGTTCGGAAGGAAGATG	135
	Reverse	AAGCCCCTTTCCACAAAAGTA	
CEP164	Forward	AACACTATCGGAGCTTGGTGAT	126
	Reverse	GGGGTCTCTGTCCTTCTTGTCT	
CNOT1	Forward	CTGTCAAACATCGTGTTGAGAAT	128
	Reverse	CCATATTTGTTTGATGTCTTCCA	
KCNMA1	Forward	CCTCTCCTTACCTCATCAGCTT	121
	Reverse	CCGTCAACACTATCACCAAAAA	
ANUBL1	Forward	CCAGCATTTTCAGGAAGAAAAC	119
	Reverse	ACTACTAGAGCATGTTGAAAGGAAAA	
EPRS	Forward	ACATCAGGGTCAAAGGAAAAGA	120
	Reverse	ACCCTGGGATAGAAAGACCATT	
TTC3	Forward	TCAGCCTAGAGAAACTAAGACTGAAA	115
	Reverse	TCCTCCTCCATTCTTTCTTGTG	
SFRS12IP1	Forward	CAGGAATAAATGAAGAAGAGGAAA	134
	Reverse	CTCATAGGCCTGCTTTGTTTTT	
SFR	Forward	TGAAGCAAAATGTTTCCTTTTTA	125
	Reverse	GCTGGTATTTTGTGAGGCTTTC	
DAB2IP	Forward	GCGCAGTTGTTAGAAGACGAG	124
	Reverse	CCAGGGTCCAGTTTTACCTTTT	
PHACTR4	Forward	AAGAGCAAGTTCTCAGGCTTTG	117
	Reverse	CCCTTAATATCTCACCTTCTGAAGTC	
MLL3	Forward	TCAGTGTGTATCTGTTGAACCAAA	133
	Reverse	ACTTCCGTTTTTACCTCATTGG	
PDS5A	Forward	CAATGCAGATTCACCAAAGGAC	94
	Reverse	TGTGGTAGGAAAAATGAGAGGAA	
TROVE2	Forward	TTTGCCTTTTTGTTAGGTTTCC	166
	Reverse	AGAACCGGTGTAGTCGATTCAT	
USP42	Forward	CTGACCTCCACAGACACAAAAA	116
		ACCCTGGGTAAGTGCAGTTCT	

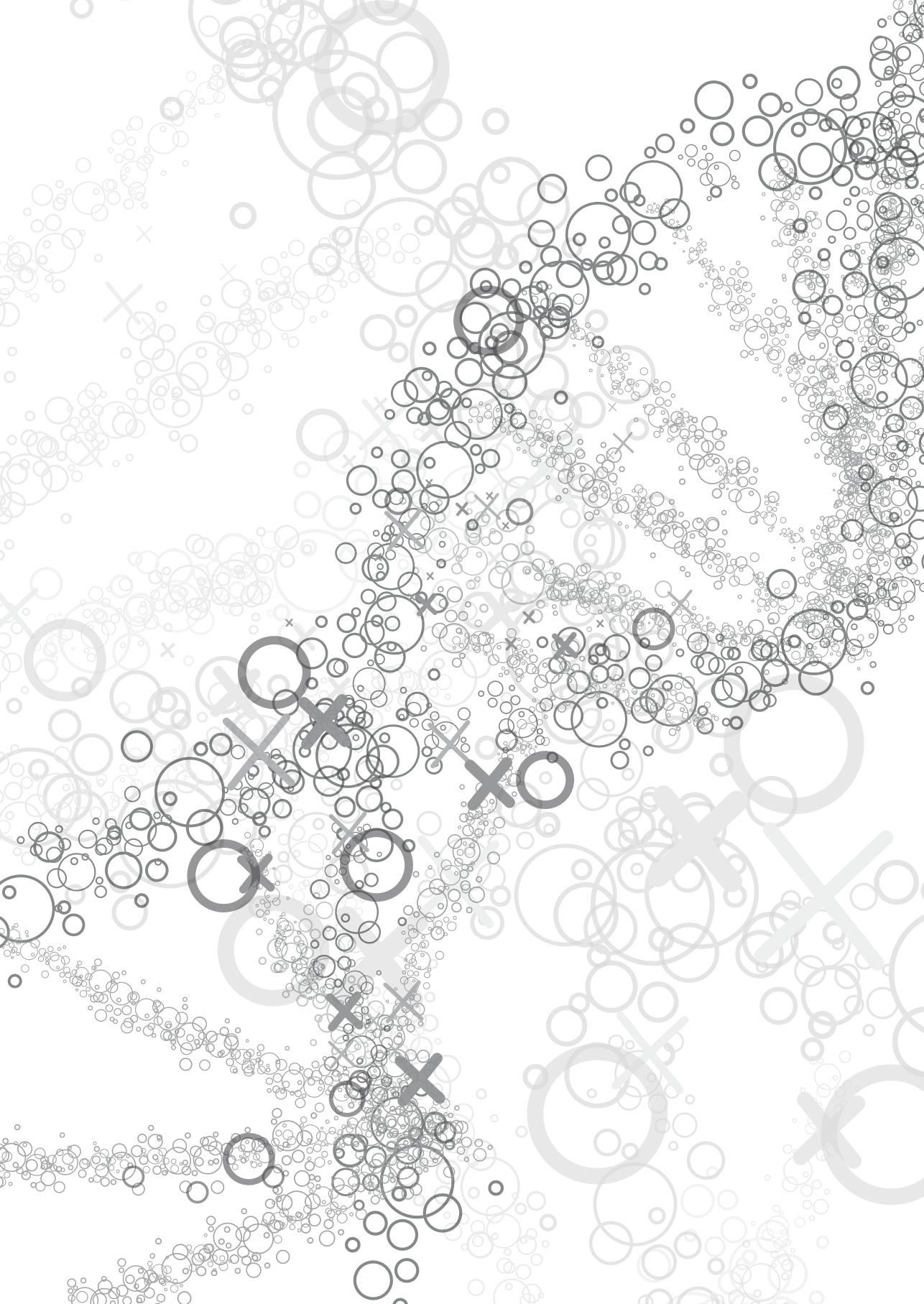
1. Carnevali P, et al., *Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads*. Journal of Computational Biology, 2012. 19(3): p. 279-292.
2. Boormans JL, et al., Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer, Int J Cancer. 2013; 133(2):335–345
3. Xi L, et al., Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. Nucleic Acids Res. 2008; 36(20):6535–6547

Supplementary Table 1 and Supplementary Table 2 will be made available through the journal's website upon publication.

Before publication they will be available through the following links:

Supplementary Table 1 -
https://www.dropbox.com/sh/qe2lx9obl3q4mel/AADhlj0J-Ox-B5XWUwz7c2caa?dl=0

Supplementary Table 2 -
https://www.dropbox.com/sh/qe2lx9obl3q4mel/AADhlj0J-Ox-B5XWUwz7c2caa?dl=0



Chapter 5

Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line

I. Teles Alves[1,2]
Saskia Hiltemann[3]
Thomas Hartjes[1]
Peter van der Spek[3]
Andrew Stubbs[3]
J. Trapman[2]
G. Jenster[1]

Departments of Urology[1], Pathology[2] and Bioinformatics[3], Erasmus MC, Rotterdam, The Netherlands

Human Genetics, 2013, 132(6):709-13

Abstract

The VCaP cell line is widely used in prostate cancer research as it is a unique model to study castrate resistant disease expressing high levels of the wild type androgen receptor (AR) and the *TMPRSS2-ERG* fusion transcript. Using next generation sequencing we assembled the structural variations in VCaP genomic DNA and observed a massive number of genomic rearrangements along the q arm of chromosome 5, characteristic of chromothripsis. Chromothripsis is a recently recognized phenomenon characterized by extensive chromosomal shattering in a single catastrophe event, mainly detected in cancer cells. Various structural events identified on chromosome 5q of VCaP resulted in gene fusions. Out of the 18 gene fusion candidates tested, 15 were confirmed on genomic level. In our set of gene fusions only rarely we observe microhomology flanking the breakpoints. On RNA level, only 5 transcripts were detected and *NDUFAF2-MAST4* was the only resulting in an in-frame fusion transcript. Our data indicate that, although a marker of genomic instability, chromothripsis might lead to only a limited number of functionally relevant fusion genes.

Advances in DNA sequencing technologies have allowed the detailed analysis of genomic aberrations in cancer. Stephens *et al.* (2011) described massive genomic rearrangements, designated 'chromothripsis', in a subgroup of chronic lymphocytic leukemia (1). Subsequently, chromothripsis has been observed in several cancer cell lines, including lung, sarcoma, esophageal, renal, thyroid cells and in a few patient samples of multiple myeloma, colorectal cancer, medulloblastoma and neuroblastoma (2-5).

The process of chromothripsis involves the shattering of one or a few chromosomes that is followed by fragment reassembly into derivative chromosomes (6). It is defined on the basis of three main features: the occurrence of numerous genomic rearrangements in localized chromosomal regions, several copy number changes alternating between only one, two, or occasionally three, different copy number states and the alternation between regions where heterozigosity is preserved with regions displaying loss of heterozigosity (6). The localized pattern observed for chromothripsis differs from that of other types of genomic instability where rearrangements tend to be dispersed genome-wide (11-12) and suggests chromothripsis will likely occur when chromosomes are largely condensed. Moreover, the alternation between few copy number states strongly implies that the chromosomal rearrangements occurred in a short time scale, probably in one single mutational event (1, 13).

Several mechanisms have been proposed to induce the massive number of genomic rearrangements observed in chromothripsis. Overall, the clustering pattern of rearrangements observed in chromothripsis is readily explained by assuming a condensed configuration of the chromosome by the time chromothripsis was triggered. Moreover, one can also consider, although less reasonably, that a chromosomal region is exposed to localized high-energy ionizing radiation (7). Along with telomere erosion and breakage–fusion–bridge cycles also abortive apoptosis after extensive chromosomal fragmentation and replication stress have been proposed as potential initiating triggers of chromothripsis (8). Currently, the most attractive model for chromothripsis combines both replication stress and mitotic errors with the formation of micronuclei containing missegregated anaphase chromosome(s) that undergo defective DNA replication.

The fragmentation created by this first stage of chromothripsis is further repaired by one of the several DNA repair mechanism (9). So far, both non-replicative repair pathways such as non-homologous end joining (NHEJ) and replication-associated repair pathways such as microhomology-mediated break-induced replication (MMBIR) have been implicated in the reassembly process of pulverized chromosome(s) (10). In case of limited sequence overlap, non-homologous end joining has been suggested as the most probable molecular mechanism involved in reconnecting the shattered DNA fragments (4).

In this study, whole-genome paired-end sequencing was performed on DNA from the prostate cancer cell line VCaP (14). This cell line is widely used as a model for research on castration-resistant prostate cancer (CRPC). It is derived from a vertebral metastatic

lesion, grows androgen-dependently and has an amplified androgen receptor (AR) gene and the common *TMPRSS2-ERG* fusion gene (15). The genome was sequenced at an average coverage of 113x and an average fully called genome fraction of 97.6% (Supplementary methods). Overall, we detected 2414 high confidence structural variation (SV) breakpoint events of which less than 4% juxtaposed portions of gene sequences on both sides (Figure 1a). By far the highest number (573; 24%) mapped on chromosome 5q (Figure 1b, Table S1).

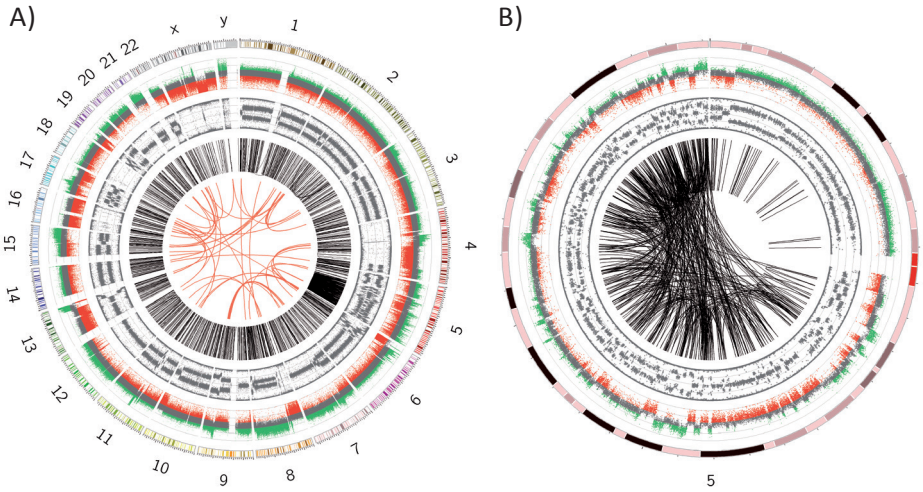


Figure 1: Circos plots showing structural and copy number variation across the whole genome (A) and chromosome 5 (B) of VCaP. Each chromosome is represented in the outer ring. The outer data ring corresponds to copy number variation, with regions of gain depicted in green and loss in red. The inner data ring represents B-allele frequency. The intra and interchromosomal rearrangements are on the inside and depicted in black and red lines, respectively.

The pattern of rearrangements observed in 5q strongly resembles the chromothripsis phenomenon recently described in cancer cells (16). First, we observed a large number of rearrangements in a localized chromosome location (5q). Second, the copy number distribution in 5q varied predominantly between two copy number states with sporadic regions of three (Figure 2, Figure S1). The regions with lower copy number have retained heterozygosity (Figure 2, Figure S1). The rare presence of loss of heterozigosity regions correlates with the fact that the VCaP cell line has a near-triploid genome, in which, most likely the chromothripsis event happened after the duplication. In addition, the breakpoints in 5q were remarkably clustered in smaller regions, with six rearrangements involving a 7 kb region between 81.121 Mb and 81.128 Mb (Table S1) although the normal distance between the two joined fragments is normally of tens of megabases.

Out of the 573 SV breakpoints involving 5q, only four were interchromosomal, involving hence rearrangements between chromosome 5 and other chromosomes

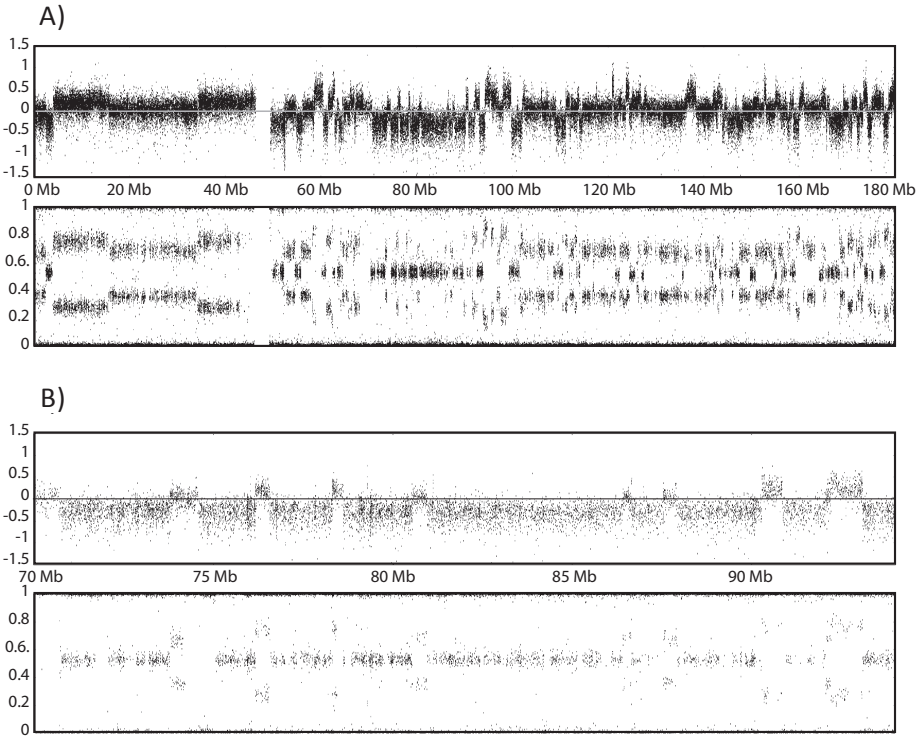


Figure 2: Clustered rearrangements on chromosome 5q of VCaP. Copy number across chromosome 5 oscillates between a copy number of 2, 3 and 4. Copy number 2 corresponds to segments of SNP probes bellow the zero line, copy number 2 to segments of SNP probes in the zero line and copy number 4 to segments of SNP probes above the zero line. The B-allele frequency plot is displayed bellow the copy number plot.

(Table S1). Remarkably, sequencing of the breakpoint junctions of these SV breakpoints revealed frequent insertions of novel sequences (186/573) (Table S2). In the gene fusions resulting from complex chromothripsis rearrangements in 5q, these insertions of up to 255 bp corresponded to fragments of chromosome 5 located within 7 to 53 Mb distance of the adjacent fusion partner. We rarely observe microhomology flanking the breakpoints in our set of gene fusions and, in a few cases, we could detect repeats flanking both sides of the breakpoint junction (Table S3).

Forty-three out of the 573 SV breakpoint events on 5q were in different genes at both sides, with 18 of these involving genes in the same orientation capable of originating functional gene fusions (Table 1, Figure S2).

In order to confirm the 18 candidate gene fusions detected by sequencing we have designed PCR reactions spanning sequences on both genes. We could validate 15

candidate gene fusions at the DNA level (Figure S3) whereas only one third were detected on mRNA level suggesting downregulation of gene expression or instability of the fusion transcripts (1). The two non-validated gene fusions *PDE4D-PPP2R2B* and *PPP2R2B-FAM172A* represent by-products of the *PDE4D-FAM172A* gene fusion in which a small fragment of *PPP2R2B* sequence is inserted in-between the *PDE4D* and *FAM172A* genes. Conversely, the *ADAMTS12-PXDNL* candidate gene fusion had a very low number of confirming DNA reads and hence is most likely a sequencing artifact (Table 1). Chromothripsis, being a seemingly random process, originates highly derivative chromosomes with numerous mutations and rearrangements. The formation of gene fusions as a consequence of the chromothripsis event does not seem to be preferential over rearrangements occurring outside genes (43 versus 573). Moreover, we did not observe positive selection for in-frame fusion transcripts since only one out of the five expressed fusion transcripts resulted in a feasible fusion protein. The role of this fusion between *NDUFAF2* and *MAST4* remains to be determined (Figure S4, Table S4). Using PCR, we observed that all 15 fusions detected in VCaP were also present in the DuCaP cell line, which is derived from a dura mater metastasis from the same patient that gave rise to VCaP (17)), indicating that chromothripsis occurred in the cells that resulted in both VCaP and DuCaP metastases (Figure S3 and S4) (data not shown). The confirmation that both VCaP and DuCaP harbor the gene rearrangements identified in chromosome 5 supports

Table 1: List of gene fusions involving chromosome 5 of VCaP. The column 5' Chr refers to the chromosome number of the 5' donor gene and the column 3' Chr refers to the chromosome number of the 3' acceptor gene. The column # of discordant mate pairs displays the number of mate pair reads that are discordant in relation to the reference genome build and concordant with the respective reported SV event (gene fusion).

5' Donor gene	3' Acceptor gene	5' Chr	3' Chr	Validated cDNA	Validated DNA	In frame	# of Discordant mate pairs
PDE4D	C5orf47	5	5	yes	yes	no	218
CPLX2	UBXD8(FAF2)	5	5	no	yes	no	55
EBF1	FBXL17	5	5	no	yes	no	74
KCNN2	EBF1	5	5	no	yes	yes	509
RASGRF2	RNF145	5	5	no	yes	no	71
JMY	DMGDH	5	5	yes	yes	no	102
TRIM40	FBXO38	6	5	no	yes	no	89
LMAN2	AP3S1	5	5	yes	yes	no	191
EFNA5	PCDHB7	5	5	no	yes	no	11
YTHDC2	PPP2R2B	5	5	no	yes	no	8
PDE8B	UIMC1	5	5	no	yes	no	63
ZFP62	RGNEF	5	5	no	yes	no	225
NDUFAF2	MAST4	5	5	yes	yes	yes	197
ADAMTS12	PXDNL	8	5	no	no	no	3
EBF1	FEM1C	5	5	no	yes	no	11
PDE4D	FAM172A	5	5	yes	yes	no	119
PDE4D	PPP2R2B	5	5	no	no	no	12
PPP2R2B	FAM172A	5	5	no	no	yes	7

the assumption that these were present in the precursor prostate adenocarcinoma lesion rather than being cell line cultivation artifacts.

In our whole genome study of VCaP cells we also detected a fairly complex rearrangement involving the *TMPRSS2* and *ERG* genes on chromosome 21q. The *TMPRSS2-ERG* fusion in VCaP results from the assembly of the *ERG* and TMPRSS2 breakpoints with the insertion of two fragments of *TMPRSS2* (Table S5). This did not disrupt the transcript and open reading frame of the final fusion product (18).

Recently, an association between *TP53* mutations and chromothripsis has been observed for acute myeloid leukemia and pediatric medulloblastoma (4). In order to determine the mutational status of *TP53* in the VCaP cell line we have examined all the single nucleotide variants (SNVs) detected by next-generation sequencing in the *TP53* gene (Table S6). We observed two homozygous missense SNVs present in the *TP53* coding DNA sequence (CDS): c.742C>T and c215C>G. The c.742C>T (p.R248W) is a well known hotspot *TP53* mutant with no residual activity whereas the c215C>G (p.R72P) is a natural occurring variation in exon 4 of *TP53*. As a result of the c.742C>T SNV the VCaP cell line has no wild type functional p53, a key player in the maintenance of genomic stability. In addition, we have also examined whether there were SNVs present in DNA repair genes previously shown to be altered in prostate cancer (19) (Table S7).

Interestingly, we found the genes *ATM, MLH1, PRKDC* and *ERCC5* to have missense SNVs in the respective CDS. The only SNVs described to be mutational were present both in the *ATM* gene (COSMIC mutation ID number 21826 and 21827) (20). An SNV present in the intron 7 acceptor splicing site of the *XRCC4* gene (rs1805377) has shown to be significantly associated with increased prostate cancer risk (21). The homozygous missense c.655A>G SNV present in the *MLH1* gene (p.I219V) has been associated with increased risk of colorectal cancer (22). Deficiencies in the DNA mismatch repair system are commonly observed in colorectal cancer and to a less extend in prostate cancer. The influence of the c.655A>G (p.I219V) missense SNV in the development of prostate cancer remains undetermined.

Here we report the structural variations detected in VCaP and show that the q arm of chromosome 5 has undergone chromothripsis. Chromosome 5 appears to be frequently affected by chromothripsis with studies showing the same pattern in a renal cancer cell line (1) and a neuroblastoma patient (5). It remains undetermined, however, whether chromosome-dependent properties or the nonrandom radial localization of chromosomes in distinct territories play a role in the preferential target of certain chromosomes by chromothripsis. Whether the catastrophic events on 5q in VCaP, and the related cell line DuCaP, have been playing and still play an important role in tumorigenesis remains to be determined.

The massive number of genomic breaks occurring in chromosome 5q hampers the formulation of a definitive model for the generation of chromothripsis. The copy number variation and B-allele frequency of chromosome 5q in VCaP is a repetition of regions

with n2 (AB), n3 (AAB) and n4 (AAAB) (Figure S1). Based on chromosome painting, VCaP is an overall near-triploid cell line with 4 copies of chromosome 5 that vary in size (23). We hypothesize that this pattern is explained by two normal chromosome copies (n2 (AB)) and two copies of A with chromothripsis (n4 (A*A*AB)) of which one of the two copies has undergone additional rearrangements resulting in n3 (A*AB). The sequence of events that would lead to this pattern is a duplication of A and B (AABB) and subsequent loss of one B allele (AAB). One of the A alleles would have undergone chromothripsis (A*AB) and duplication (A*A*AB). Additional rearrangements (deletions or even a second chromothripsis event) would explain the observed 5q regions of n3 A*AB (Figure S5). Our study has shown that research on genes located on and transcripts derived from chromosome 5q, need to take the chromothripsis into account. Besides being a widely used model for research on AR, ERG and CRPC, VCaP might prove a highly relevant model for research on chromothripsis.

References

1. Stephens PJ, Greenman CD, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011;144:27-40
2. Magrangeas F, Avet-Loiseau H, Munshi NC, Minvielle S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. Blood 2011;118:675-678
3. Kloosterman WP, Hoogstraat M, Paling O, et al. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biology 2011;12
4. Rausch T, Jones DT, Zapatka M, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 2012;148:59-71
5. Molenaar JJ, Koster J, Zwijnenburg DA, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature 2012;483:589-593
6. Maher CA, Wilson RK. Chromothripsis and human disease: piecing together the shattering process. Cell 2012;148:29-32
7. Misteli T. Beyond the Sequence: Cellular Organization of Genome Function. Cell 2007;128:787-800
8. Jones Mathew JK, Jallepalli Prasad V. Chromothripsis: Chromosomes in Crisis. Developmental Cell 2012;23:908-917
9. Holland AJ, Cleveland DW. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat Med 2012;18:1630-1638
10. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. 2012;12:663-670
11. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature Genetics 2008;40:722-729
12. Stephens PJ, McBride DJ, Lin M-L, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes.
13. Righolt C, Mai S. Shattered and stitched chromosomes-chromothripsis and chromoanasynthesis-manifestations of a new chromosome crisis? Genes Chromosomes Cancer 2012;51:975-981
14. Drmanac R, Sparks AB, Callow MJ, et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 2010;327:78-81
15. Korenchuk S, Lehr JE, L MC, et al. VCaP, a cell-based model system of human prostate cancer. In Vivo 2001;15:163-168
16. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer 2012
17. Lee YG, Korenchuk S, Lehr J, et al. Establishment and characterization of a new human prostatic cancer cell line: DuCaP. In Vivo 2001;15:157-162
18. Mertz KD, Setlur SR, Dhanasekaran SM, et al. Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model. Neoplasia 2007;9:200-206
19. Grasso CS, Wu Y-M, Robinson DR, et al. The mutational landscape of lethal castration-resistant prostate cancer. 2012;487:239-243

20. Gumy-Pause F, Wacker P, Maillet P, Betts DR, Sappino A-P. ATM alterations in childhood non-Hodgkin lymphoma. Cancer Genetics and Cytogenetics 2006;166:101-111
21. Mandal RK, Singh V, Kapoor R, Mittal RD. Do polymorphisms in XRCC4 influence prostate cancer susceptibility in North Indian population? Biomarkers 2011;16:236-242
22. Campbell PT, Curtin K, Ulrich CM, et al. Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. Gut 2009;58:661-667
23. van Bokhoven A, Caires A, Maria MD, et al. Spectral karyotype (SKY) analysis of human prostate carcinoma cell lines. Prostate 2003;57:226-244

Supplementary data:

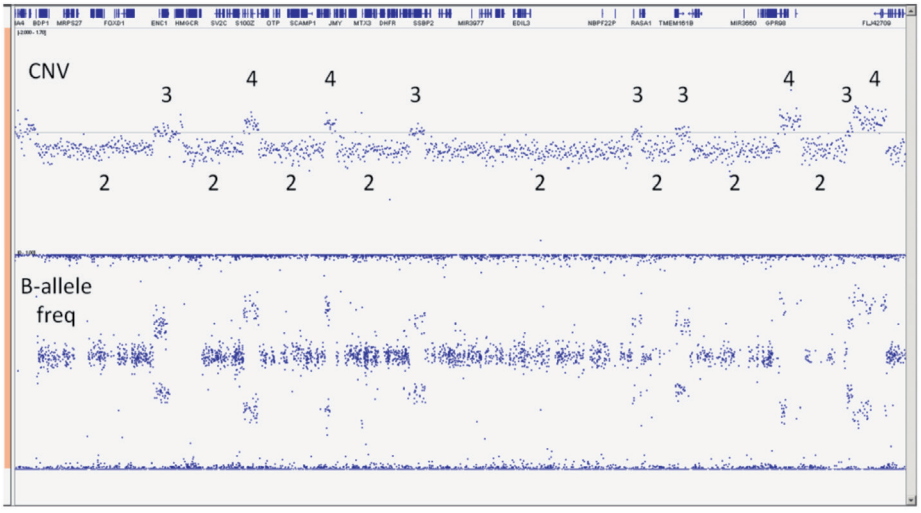


Figure S1: Copy number and B-allele frequency distribution across a region on chromosome 5q. Each blue dot represents a SNP detected on the 1M SNP Illumina array. Chromosome 5 is depicted on top with the red box corresponding to the zoomed-in region. The copy number variation plotted varies from copy number state 2 to 4.

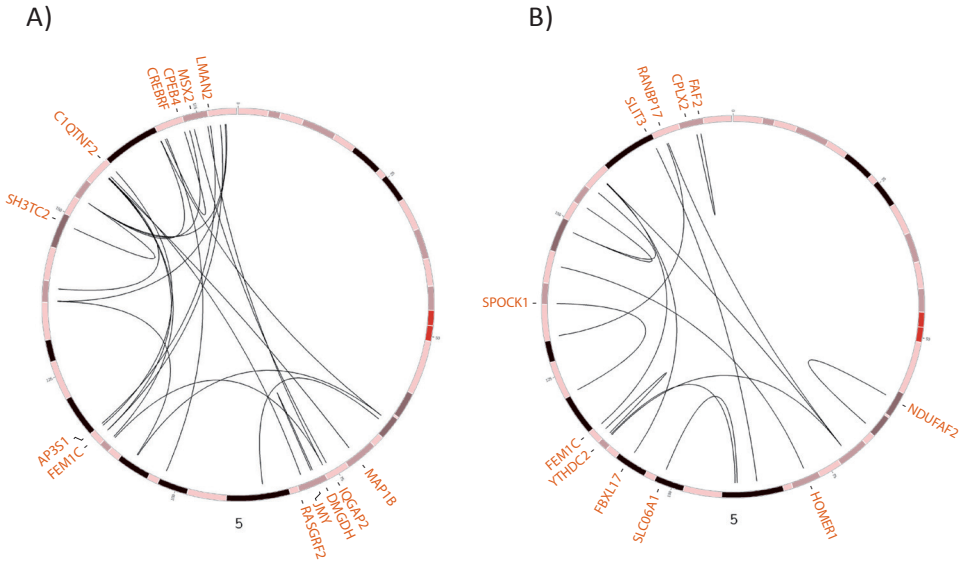


Figure S2: Circos plots of the 43 gene rearrangements present on chromosome 5q of VCaP. (A) represents 25 rearrangements between genes of different orientations and (B) represents 18 rearrangements between genes with the same orientation.

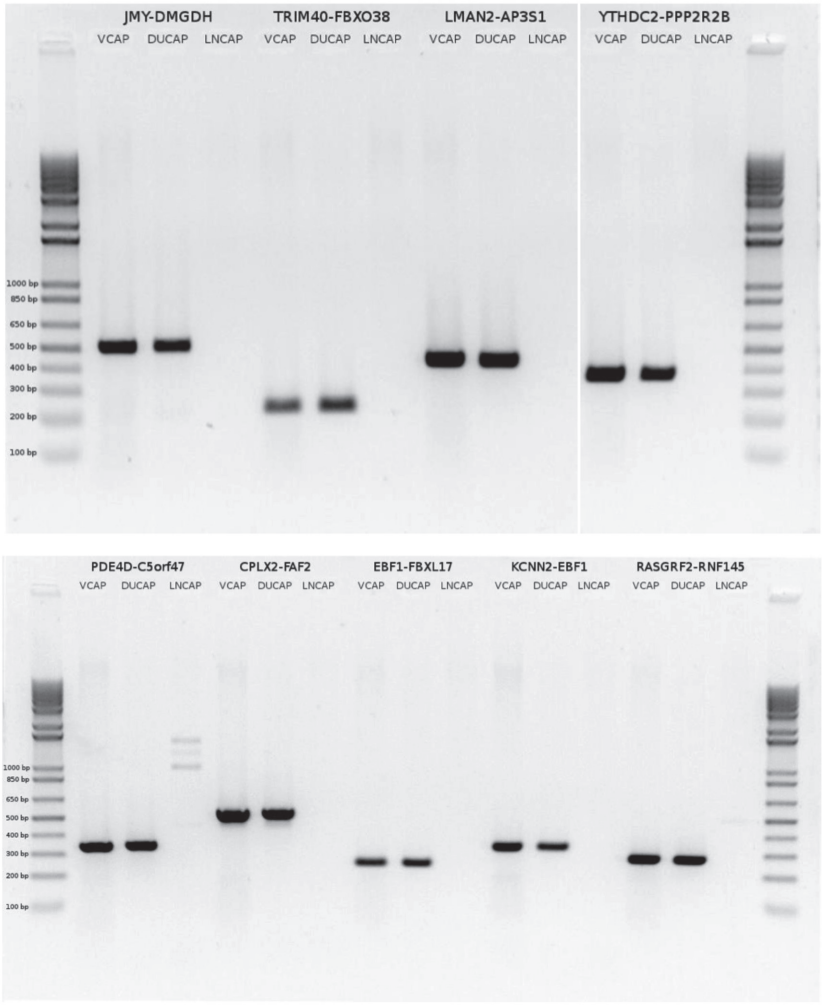


Figure S3: RT-PCR of the validated DNA fusions (electrophoresis on a 1% agarose gel). The gene fusion name and sample is displayed on top of the corresponding wells. We have used VCaP and DuCaP as our test samples and LNCaP as a control for specificity. LNCaP is a cell line derived from a lymph node carcinoma of the prostate. It is nearly-tetraploid and displays a stable karyotype [1-2]. The positive control consists of primers within the NCAPD2 gene.

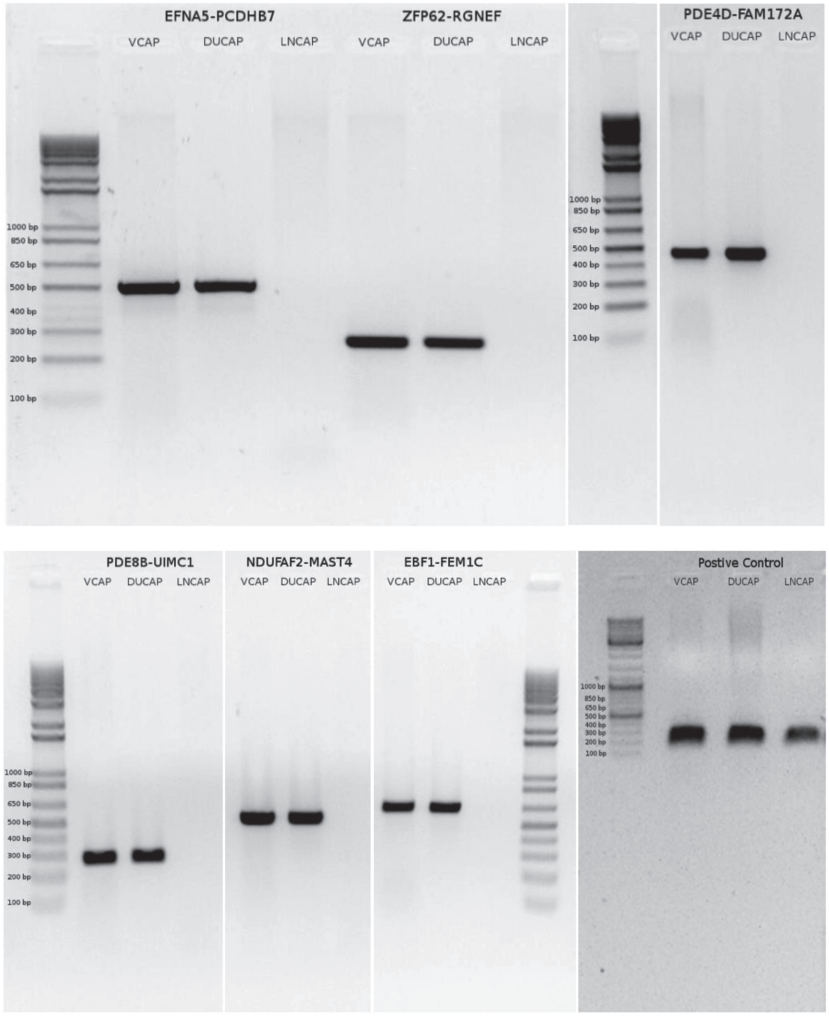


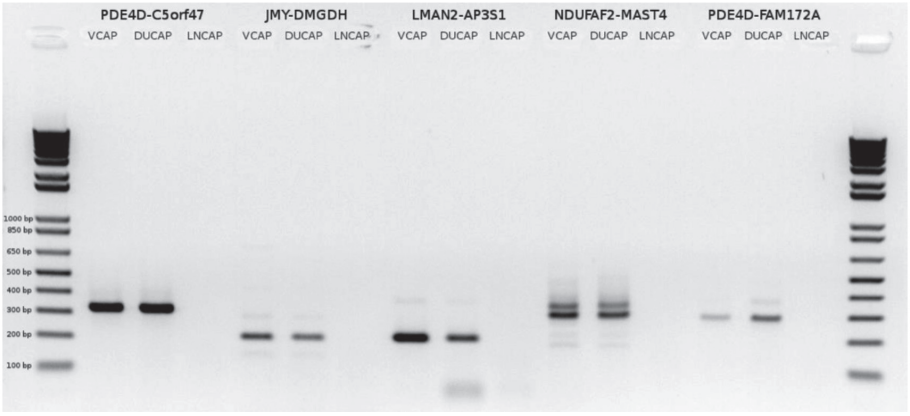
Figure S3: Continued

Table S2: DNA alignment of the validated gene fusions. The DNA sequences corresponding to the first 30 bp before and after the junctions obtained by sequencing of the PCR products. The red nucleotide here present represent a difference in relation to the reference genome. The transition sequence found in the junction is depicted in red in an additional row and the red nucleotides in the 30bp sequences represent variation found in the sequencing of the PCR products to the reference genome.

PDE4D	[ENST00000405755]	Intron 5-6	-	C5orf47	[ENST00000522195]	Intron 1-2
AATGACAGTAATTGGAAGCAGTGTAGAAAG				TCTCTGTTTTCCTGGTATGTTCCTGCAGTA		
				C		
CPLX2	[ENST00000359546]	Intron 2-3	-	UBXD8	[ENST00000261942]	Intron 10-11
CAGGAGAATCGCTTGAACCTGGGAGGCGGA				GCTTGTGGTGAGCTGAGATGGCGCCATTGC		
FBXL17	[ENST00000359660]	Intron 5-6	-	EBF1	[ENST00000313708]	Intron 11-12
TATTTAACATACAAATAGTAACATCATCAA				TGGTATGCCTTCTTCATTAACTGCTCATAA		
KCNN2	[ENST00000264773]	Intron 5-6	-	EBF1	[ENST00000313708]	Intron 11-12
TCAAGAGACAAAATCACCTGCAGGAGAGAA				TTCAAGACAGAGTCATATATGTCACCAATA		
RASGRF2	[ENST00000265080]	Intron 2-3	-	RNF145	[ENST00000274542]	Intron 2-3
ACTGTATTCATTTTCTATTCTGTAACAAAT				GGGAAAGTTCCTTTATGTTGCAAAACACTG		
DMGDH	[ENST00000255189]	Intron 7-8	-	JMY	[ENST00000396137]	Intron 2-3
GCCCTTCTGTGTTCCAGGAAATAGTTGCCC				CGTAATTAGTCGATTGTACCTAAAACCTTA		
FBXO38	[ENST00000340253]	Intron 12-13	-	TRIM40	[ENST00000383610]	Intron 2-3
TGGGTGGTGATGCCCCTGGGCTAGTGAAGA				AAATTTAACGTGGAGTTATCCTTCACACAA		
AP3S1	[ENST00000316788]	Intron 1-2	-	LMAN2	[ENST00000303127]	Exon 2
GTCTCTACTAAAAATACAAAAATAGCTGGG				AGACGTACGTACTGGCTCGTGAGCATAGTG		
AGTGAGATAGCAGGGTAAGGTTAACCCTCACCTCCCCAGCACAGAAAGGACTG						
CAGAAAGCCCCGAGAGCCTCAGGCTGCCTCCACCTCTGTCTCTGCAGTGCCTTT						
EFNA5	[ENST00000333274]	Intron 1-2	-	PCDHB7	[ENST00000231137]	Exon 1
GGTTCCTGCCAGTGGGTATGCGACAGGAAG				GCTGCAGAATGGCTCCGCGCCCTGCACCGA		
YTHDC2	[ENST00000161863]	Intron 2-3	-	PPP2R2B	[ENST00000394413]	Intron 5-6
CTTACTCTATGTCCCCGCAAAGGACATGAA				TTGAAGCCCAGAGAGTGAAGTGGAGCATGG		
AATTTGATTGTGCTGTGGTCTGAGAGAGTGGTTGTTGTAATATCAATTCTTT						
TGCATTTGCTGAGCAAAGTTTTGTATTTGATTATATAATCAGTTTTAGAGTA						
PDE8B	[ENST00000264917]	Intron 1-2	-	UIMC1	[ENST00000274827]	Intron 10-11
CTTTTTTTGGTTGCTCCAGAGAAATTGTGA				CAGAATGAATACTGGTAACACTCCAACTTG		
NDUFAF2	[ENST00000296597]	Intron 1-2	-	MAST4	[ENST00000403625]	Intron 2-3
AAGGCTGCCCAAATATCTCTCCTGGCCTGT				TGTGCAGGGTTGAATAACAATGGTTCTTAA		
FEM1C	[ENST00000274457]	Intron 2-3	-	EBF1	[ENST00000313708]	Intron 6-7
GTAACAAGACAACTCACAAAAGACTACATA				ACATTGCTCTATGACTGCCACTTACTAATC		
GGGACATGAACATCAGAATGACTTGATAACATTGCTTGTTTTGAGCTTTTCCTTCAAAT						
TTTAGCCCCTGATGTGGTTTGTTTACGGTAACAACAACAGGAATACTCACTGAGCCACT						
ACTACATACAGGCACTTTGCTCAGTGCTGGAAATACAGCAGTGATCCATGGACCAAAA						
TCCCTGTTCTAGTGCAGATGACATTCTTTCTAATGAGGTGAGAAAGATAATAAGCAACA						
CAAATAATAAATAAGTAAGT						
PDE4D	[ENST00000405755]	Intron 1-2	-	FAM172A	[ENST00000265139]	Intron 9-10
ATTTATGTGCTATGTACTGTGTAATCCACC				AGAATTGTCAGCCACTAAGTTTATCAAGCT		
CAGCATGCTTCAGTGGTACCACTATGAGTCTTTAGACCTGGGGAAACTACACCTT						

Table S3: List of repeats flanking both sides of the junction

5' Donor gene	3' Acceptor gene	5' Chr	3' Chr	LeftRepeat Classification	RightRepeat Classification
PDE4D	C5orf47	5	5		L1PB1:LINE:L1
CPLX2	UBXD8	5	5	AluY:SINE:Alu	AluY:SINE:Alu;Self chain
EBF1	FBXL17	5	5		
KCNN2	EBF1	5	5	MER20:DNA: MER1_type	
RASGRF2	RNF145	5	5	MLT1E2:LTR:MaLR	
JMY	DMGDH	5	5		
TRIM40	FBXO38	6	5		
LMAN2	AP3S1	5	5	L2:LINE:L2	L1PA8:LINE:L1
EFNA5	PCDHB7	5	5	AluSc:SINE:Alu;Self chain	Self chain
YTHDC2	PPP2R2B	5	5		SegDup;Self chain
PDE8B	UIMC1	5	5		AluSx:SINE:Alu; Self chain
ZFP62	RGNEF	5	5		
NDUFAF2	MAST4	5	5		LTR71A:LTR:ERV1
EBF1	FEM1C	5	5		L1MB8:LINE:L1
PDE4D	FAM172A	5	5		



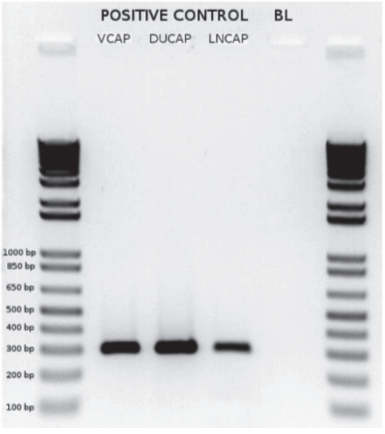


Figure S4: RT-PCR of the validated RNA fusions (electrophoresis on a 1% agarose gel). The gene fusion name and sample is displayed on top of the corresponding wells. We have used VCaP and DuCaP as our test samples and LNCaP as a control for specificity. BL represents the blank control (no cDNA). The positive control consists of primers within the GAPDH gene.

Table S4: RNA alignment of the validated gene fusions. The sequencing data was used to predict the gene fusions at RNA level. The RNA sequences and the corresponding protein translations of the first 30 bp before and after the junctions obtained by sequencing of the PCR products are depicted below. * represents the fusions in frame.

PDE4D	[ENST00000405755]	Exon 5	-	C5orf47	[ENST00000522195]	Exon 2	
AACCATCCATCAACAAAGCCACCATAACAG				GCAGCAAGCCAATCACTGAGGTGATGCGTT			
Q--P--S--I--N--K--A--T--I--T--				.....			
JMY	[ENST00000396137]	Exon 2	-	DMGDH	[ENST00000255189]	Exon 8	
CTTGCCATGCTACGAAGACAGCAGATCAAG				ATATGGCATAATCCACGCTGGTGGGGTAGG			
-L--A--M--L--R--R--Q--Q--I--K-				--Y--G--I--I--H--A--G--G--V--G			
LMAN2	[ENST00000303127]	Exon 1	-	AP3S1	[ENST00000316788]	Exon 2	
GGGAGCATTCGCTCATTAAGCCCTACCAAG				AGTGAAGATACACAACAGCAAATCATCAGG			
R--E--H--S--L--I-=K=-P--Y--Q--				-S--E--D--T--Q--Q--Q--I--I--R-			
* NDUFAF2	[ENST00000296597]	Exon 1	-	MAST4	[ENST00000403625]	Exon 3	
ACTACATCCCGCAGTACAAGAACTGGAGAG				GGAGGTACCTTCTTCCAAACCCGGTGGCGG			
Y--Y--I--P--Q--Y--K--N--W--R--				G--R--Y--L--L--P--N--P--V--A--			
PDE4D	[ENST00000405755]	Exon 1	-	FAM172A	[ENST00000265139]	Exon 10	
GATTGCTATCACTTCTGCAGAATCCAGTGG				GCACCGACCGTCACGAGCTAACTTCCTGGA			
--I--A--I--T--S--A--E--S--S--G				G--T--D--R--H--E--L--T--S--W--			

Table S5: List of junction related to the event generating the TMPRSS2-ERG gene fusion in VCaP. The scheme bellow represents the three junctions with matching position numbers and leading to the TMPRSS2-ERG fusion.

Left gene	Chr	Position	seqO	Righ Gene	Chr	Position	seqO	Mate pairs
TMPRSS2	chr21	41779386	-	TMPRSS2	chr21	41793822	+	307
\	chr16	66265351	-	ERG	chr21	38798283	+	12
TMPRSS2	chr21	41775625	-	TMPRSS2	chr21	41786403	-	13
C16orf70	chr16	65702685	-	TMPRSS2	chr21	41775590	-	49
ERG	chr21	38798223	+	TMPRSS2	chr21	41779892	-	474
C16orf70	chr16	65702689	+	\	chr16	66265338	-	294

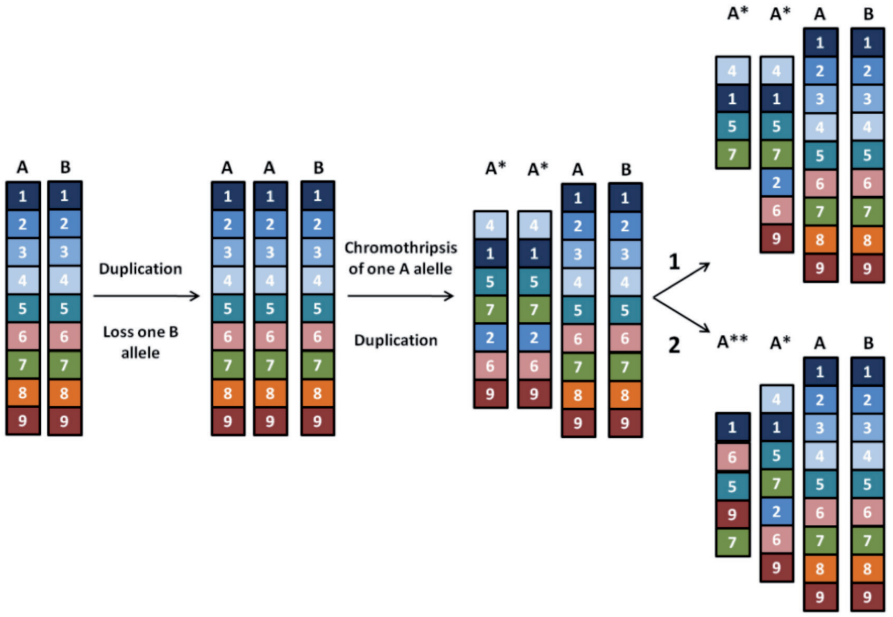


Figure S5: Hypothesis on the chain of events leading to copy number variation and B- allele frequency profile of VCaP. As can be observed in figure S1, the most likely sequence of events explaining VCaP's copy number and B-allele frequency will be a duplication of the two normal chromosome copies followed by the loss of one B allele (although the duplication of only A cannot be excluded). Next, one of the A alleles undergoes chromothripsis and is duplicated. Now, with two copies of A*, one wild type copy of A and one wild type copy of B there are two possible situations: 1. One of the A* suffers additional deletions, or 2. One of the A* undergoes another chromothripsis event.

Table S6) Single nucleotide variations (SNVs) detected by next-generation sequencing in the TP53 gene of the VCaP cell line. All except intronic SNVs are shown. The column type refers to the variation type (snp – single nucleotide polymorphism; sub – substitution; del – deletion). The reference is the base present at that position in the reference genome and the call is the base present at that position in the VCaP sample. Haplotype discriminates between heterozygous (het) and homozygous (hom) SNVs. Component refers to the component of the gene that is affected (CDS – coding DNA sequence; TSS-UPSTREAM – transcription start site-upstream). Protein Pos is the aminoacid position that is modified by the SNV, Ref Seq is the aminoacid present in the reference protein sequence and Sample Seq the aminoacid present in the VCaP protein sequence.

Chr	Position	Type	Reference	Call	Haplotype	Component	Protein Pos	Ref Seq	Sample Seq
chr17	7518263	snp	G	A	hom	CDS	247	R	W
chr17	7520196	snp	G	C	hom	CDS	71	P	R
chr17	7532892	snp	C	G	hom	TSS-UPSTREAM			
chr17	7534043	snp	T	C	hom	TSS-UPSTREAM			
chr17	7534564	snp	A	G	hom	TSS-UPSTREAM			
chr17	7535187	snp	G	T	hom	TSS-UPSTREAM			
chr17	7535209	snp	T	C	hom	TSS-UPSTREAM			
chr17	7535252	del	AAACAAA CAAACAA AC		hom	TSS-UPSTREAM			
chr17	7535434	sub	G	TT	hom	TSS-UPSTREAM			
chr17	7535500	snp	C	T	hom	TSS-UPSTREAM			
chr17	7535712	snp	A	G	hom	TSS-UPSTREAM			
chr17	7536162	snp	A	G	hom	TSS-UPSTREAM			

Table S7: Single nucleotide variations (SNVs) detected by next-generation sequencing in the selected DNA repair genes in the VCaP cell line. All except intronic and TSS (transcription start site) SNVs are shown. The column type refers to the variation type (snp – single nucleotide polymorphism; ins – insertion; del – deletion). The reference is the base present at that position in the reference genome and the call is the base present at that position in the VCaP sample. Haplotype discriminates between heterozygous (het) and homozygous (hom) SNVs. Component refers to the component of the gene that is affected (CDS – coding DNA sequence; UTR3 – untranslated region 3'; UTR5 – untranslated region 5'; acceptor – acceptor splicing site and donor – donor splicing site). Protein Pos is the aminoacid position that is modified by the SNV, Ref Seq is the aminoacid present in the reference protein sequence and Sample Seq the aminoacid present in the VCaP protein sequence.

Chr	Position	Type	Reference	Call	Haplotype	Component	Protein Pos	Ref Seq	Sample Seq
ATM	107599042	snp	G	A	het	UTR5			
ATM	107643212	snp	T	C	het	CDS	857	F	L
ATM	107648665	snp	C	G	het	CDS	1053	P	R
ATM	107656917	ins		A	het	ACCEPTOR	1134	TAATTTCTTTTTAAG	TAAATTTCTTTTTAAG
ATM	107663343	snp	T	C	het	UTR5			
ATM	107688376	snp	A	G	hom	CDS	1982	N	S
ATM	107741992	snp	G	T	het	UTR3			
ATM	107743048	snp	C	T	het	UTR3			
ATM	107744837	snp	G	T	het	UTR3			
BRCA1	38449933	snp	G	A	hom	UTR3			
BRCA1	38450347	del	TT		het	UTR3			
BRCA1	38450799	snp	C	A	hom	UTR3			
BRCA1	38476619	snp	T	C	hom	CDS	1612	S	G
BRCA1	38487995	snp	A	G	hom	CDS	1435	S	S
BRCA1	38497525	snp	T	C	hom	CDS	1182	K	R
BRCA1	38497960	snp	T	C	hom	CDS	1037	E	G
BRCA1	38498461	snp	G	A	hom	CDS	870	P	L
BRCA1	38498762	snp	A	G	hom	CDS	770	L	L
BRCA1	38498991	snp	G	A	hom	CDS	693	S	S
BRCA2	31787791	snp	A	G	het	UTR5			
BRCA2	31810298	snp	T	C	het	CDS	1268	V	V
BRCA2	31811054	snp	A	G	hom	CDS	1520	L	L
BRCA2	31813004	snp	G	C	hom	CDS	2170	V	V
BRCA2	31827386	snp	T	C	hom	CDS	2465	V	A
BRCA2	31834645	snp	T	C	het	ACCEPTOR	2601	CTTTTATTTGTTCAG	CCTTTATTTGTTCAG
BRCA2	31871011	snp	A	C	het	UTR3			
BRCA2	31871275	snp	A	G	het	UTR3			
ERCC2	50560148	snp	T	G	het	CDS	155	R	R
ERCC4	13933507	snp	G	A	hom	ACCEPTOR	324	AACTTTTCGTATTAG	AACTTTTCATATTAG
ERCC5	102296198	snp	A	G	hom	UTR5			
ERCC5	102296375	snp	T	C	het	UTR5			
ERCC5	102302517	snp	T	C	hom	CDS	45	H	H
ERCC5	102312459	snp	C	T	het	CDS	319	D	D
ERCC5	102318565	snp	A	G	het	CDS	878	N	S
ERCC5	102325849	snp	G	C	hom	CDS	1052	G	R
ERCC5	102325930	snp	G	C	hom	CDS	1079	G	R
ERCC5	102326337	snp	G	A	het	UTR3			
MLH1	37028571	snp	A	G	hom	CDS	218	I	V
PRKDC	48901857	snp	G	A	het	CDS	2898	R	C
PRKDC	48903465	ins		A	het	ACCEPTOR	2798	TTGAATTTTTTACAG	TTGAATTTTTTTACAG
PRKDC	48968369	ins	G	hom	CDS	1243	F	L	
PRKDC	49009071	snp	T	C	het	DONOR	541	GTAACA	GTAACG
RAD50	131920877	snp	G	A	hom	UTR5			
XRCC4	82684698	snp	G	A	het	ACCEPTOR	297	TGATTTTCTTTTCAG	TGATTTTCTTTTCAA
XRCC4	82684732	snp	T	G	het	CDS	308	S	S
XRCC4	82685222	snp	T	C	het	UTR3			

Table S8: Primers for the validation of the fusions on DNA level.

1	PDE4D Fw DNA	TCAGGGGACTGGGAACAGTA	340
	C5orf47 Rv DNA	GGCAAGTAGCAGGCAGGACT	
2	CPLX2 Fw DNA	CAACGTAACGCAATTCATGC	541
	UBXD8 Rv DNA	CCACAACTTGGCTCTCAAAA	
3	FBXL17 Fw DNA	TGAAATCCAGAGAAGAAGATGGA	248
	EBF1 Rv DNA	AATGGGGCATTAATGGTTTG	
4	KCNN2 Fw DNA	AACCTTGGCACTCGTGACTT	327
	EBF1 Rv DNA	TCTCCCACATAAAGCCCAAG	
5	RASGRF2 Fw DNA	GTGATCGTAGCCTCGTGACA	272
	RNF145 Rv DNA	AAGCCAAGTACCCAGTTGTTC	
6	DMGDH Fw DNA	ATGCAAAGAGAAAGCTCCAG	516
	JMY Rv DNA	TGTGATTTGCACTCTCAAACCT	
7	FBXO38 Fw DNA	ATGGTGCCTTTGGGCTAGT	244
	TRIM40 Rv DNA	TGAGTGTGCTGCTCTGTTTT	
8	AP3S1 Fw DNA	TAAAGCCAAGGATTGGGTCA	319
	LMAN2 Rv DNA	TGGCCTCTTAGCACCTGTCT	
9	EFNA5 Fw DNA	ACATGACCTCGCCATAAAGG	317
	PCDHB7 Rv DNA	ACCACCAGGTAGACGGTGAG	
10	YTHDC2 Fw DNA	CTTTCCCCCTAATCCCTGAC	261
	PPP2R2B Rv DNA	CTGAAAGGTGAGGCTTCTGC	
11	PDE8B Fw DNA	ATGTGCTGAACGCTATACTTTGC	606
	UIMC1 Rv DNA	GGTCAAGTTGGAGTGTTACCAG	
12	RGNEF Fw DNA	CCCTTAGAAGCCACCATGTC	254
	ZFP62 Rv DNA	GGAAAACTAAGGGCCTCCAA	
13	NDUFAF2 Fw DNA	AGCTCGTGTCAGGTTTGCTTAT	474
	MAST4 Rv DNA	GGAACTGTGTAACCTGCTCTCC	
14	FEM1C Fw DNA	CAACCCAATAATGCCACTCC	395
	EBF1 Rv DNA	TGTTCCGGAATCAAGCAGATA	
15	PDE4D Fw DNA	CCTCTTCCCCCATAAAACG	398
	FAM172A Rv DNA	AAGCAGACTGGTGCCAAAA	

Table S9: Primers for the validation of the fusions on RNA level

1	PDE4D Fw cDNA	ATCCGACAGCGATTATGACC	160
	C5orf47 Rv cDNA	TCAGGTTTGTTCTCCCGTCT	
2	CPLX2 Fw cDNA	ATCTGCTCTGCTCAGCGACT	127
	UBXD8 Rv cDNA	TGGGAAAATTGGCTTCAATC	
3	EBF1 Fw cDNA	AGGGGACAATTTCTTTGATGG	325
	FBXL17 Rv cDNA	CGTAGGTCCAAGCTGGAAAG	
4	KCNN2 Fw cDNA	TAATTGCCGCATGGACTGT	623
	EBF1 Rv cDNA	GCTCGTGGTGACGGAGTTAT	
5	RASGFR2 Fw cDNA	CTGCAGCTGCGAACGAAC	382
	RNF145 Rv cDNA	AAGGCTGTGGTAAACCGATTC	
6	JMY Fw cDNA	GCCTTGCAGAAGCTACAACC	206
	DMGDH Rv cDNA	CATTTGCCATAGCGATTAGGA	
7	TRIM40 Fw cDNA	AACCACCAGAAGAGGGTGTG	375
	FBXO38 Rv cDNA	AACGCCGAGTACCGTTCTTA	
8	LMAN2 Fw cDNA	TTTTGTTGTTGGGGTCTGTG	210
	APS3S1 Rv cDNA	CAGTTTGTTGTCAGATCCTCCA	
9	EFNA5 Fw cDNA	CTGGTGCTCTGGATGTGTGT	263
	PCDHB7 Rv cDNA	ACCACCAGGTAGACGGTGAG	
10	YTHDC2 Fw cDNA	ATTCGCATTGATGAGGAGGT	257
	PPP2R2B Rv cDNA	ACACGAAGGTGTTGCAATGA	
11	PDE8B Fw cDNA	GTGATCTACTGCCGGGACTC	350
	UIMC1 Rv cDNA	TGTCCCACTGTCACTCTTGG	
12	ZFP62 Fw cDNA	CTGTCTTCATGGTTTCATAGCC	333
	RGNEF Rv cDNA	GTGCTGTCCCTTCGACATTT	
13	NDUFAF2 Fw cDNA	TTGTGGAGATCGCTGTCAAG	301
	MAST4 Rv cDNA	GTGACTGCCCATTGCCTATT	
14	EBF1 Fw cDNA	TGAAGGCCAAGACAAGAACC	392
	FEM1C Rv cDNA	GTATTTCAAAGCCCCAAGCA	
15	PDE4D Fw cDNA	GAACATTCAACGACCAACCA	289
	FAM172A Rv cDNA	CTCGACACAAGCAGTGAGGA	

Supplementary methods:

Paired-end sequencing of the VCaP DNA sample was performed with the Complete Genomics service provider using a proprietary sequencing-by-ligation technology and primary data analysis, including image analysis, base calling, alignment and variant calling (1). Reads were mapped to the NCBI Build 36.1 reference genome and mappings were expanded by local de novo assembly on all regions of the genome that contain single nucleotide variations (SNVs) relative to the reference genome (1). Structural variations (SV) reported in the junctions files do not contain SNVs, inserts and deletions (indels) or substitutions (subs) but include information on junctions represented by 3 or more discordant mate pairs. The "HighConfidenceJunctionBeta" files contain all junctions which have been filtered for their absence from a set of v2.0 baseline genomes (as provided at the website of Complete Genomics (B36baseline-junctions.tsv)) (2). The junctions (left and right positions) are annotated with gene features (exons, gene symbol) derived from UCSC genome browser (3) using the rtracklayer package (4) in R (3-4). The junctions located inside such exons are further categorized as junctions with different genes on both sides, with the same genes or junctions with only one gene on either the left or right side. The following table depicts how the single intrachromosomal structural variation events are called per junction. Note from personal communication with CGI: "+/+ junction may indicate a deletion or insertion, but our current process only detects deletions."

Junction Left Strand	Junction Right Strand	Single Event
+	+	deletion
-	-	inversion or deletion
+	-	inversion
-	+	inversion

Data analysis was performed on Hg18 annotated genome of VCaP. CGA tools were used to analyze the data (http://cgatools.sourceforge.net/docs/1.3.0/) (5).

The data here obtained will be deposited in the Sanger Institute Catalogue Of Somatic Mutations In Cancer (COSMIC).

Copy number variation and B-allele frequency was obtained using Infinium Illumina Human 1M probe BeadChip microarrays. We have used the values of ≤ -0.2 for copy number loss and ≥ 0.2 for copy number gain.

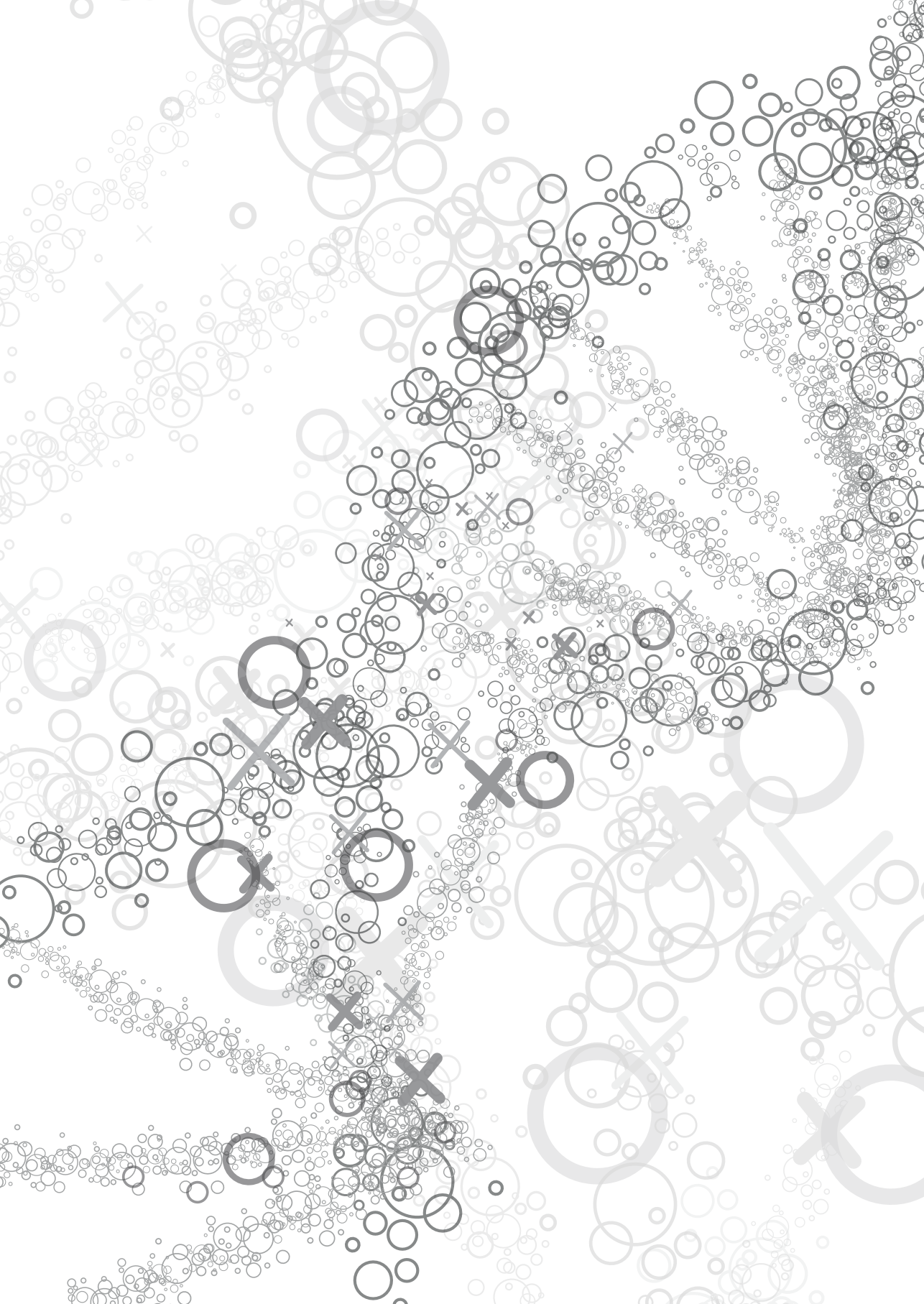
DNA and RNA isolation were performed using the RNeasy kit (Qiagen) and the QIAamp DNA Blood Midi Kit (Qiagen) according to the manufacturers protocol. Concentration and purity of DNA and RNA was assessed using the NanoDrop ND 1000 spectrophotometer (Nanodrop products) by absorption measurements at 280 nm and 260 nm, respectively. RNA was stored at -80˚C. DNA was stored at -20˚C.

VCaP was cultured in RPMI-1640 (Lonza) with 10% FCS (PAN Biotech) and 1% penicillin-streptomycin (BioWhittaker). LNCaP and DuCaP were cultured in DMEM (BioWhittaker) with 5% FCS (PAN Biotech) and 1% penicillin-streptomycin (BioWhittaker). All cell lines were cultured at 37˚C with 5% CO_2.

cDNA was synthesized using 1 µg total RNA, M-MLV reverse transcriptase kit (Promega), and Oligo (dT) primer (Invitrogen) according to manufacturer's protocol. Reverse transcription was performed at 37˚C for 60 minutes and 95˚C for 10 minutes. The PCR for multiple fusions was carried out by amplification of the cDNA or DNA samples with the HotstarTaq Kit (Qiagen). For the standard PCR reactions, 1 µL of cDNA was used in a 50 µL reaction containing 5 µL 10x PCR Buffer, 2 µL of each primer solution (100 µM), 0.25 µL HotstarTaq (5 units/µL), 1 µL of dNTPs (10 mM), and 40.75 µL of nuclease free water. An initial denaturation step of 15 minutes at 95˚C was used to activate the HotstarTaq, followed by 35 cycles consisting of a denaturation step at 95˚C for 30 seconds, an annealing step at 55˚C for 30 seconds and an elongation step at 72˚C for 1 minute. A final elongation step at 72˚C for 10 minutes was used. PCR products were checked by electrophoresis of 20 uL of product in a 1% agarose gel.

Purified PCR products have been sequenced bi-directionally using standard Sanger sequencing. A sequencing PCR was carried out using the same forward and reverse primers as used during the RT-PCR except with different primer concentrations (3 ng of primer per reaction) in a 20 µL reaction volume. The PCR product was sequenced on an ABI Model 3730 automated sequencer and analyzed using DNAMAN (Lynnon Corporation).

1. Carnevali P, Baccash J, Halpern AL, et al. Computational Techniques for Human Genome Resequencing Using Mated Gapped Reads. J Comput Biol. 2012;19:279-292
2. http://www.bx.psu.edu/~giardine/tests/tmp/DataFileFormats112.pdf
3. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841-1842
4. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Research. 2002;12:996-1006
5. Molenaar JJ, Koster J, Zwijnenburg DA, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature. 2012;483:589-593



Chapter 6

General Discussion

Prostate cancer (PCa) represents a major burden in our society. It is a disease of the elderly, a disease of the westernized world and a disease that affects different races differently. Since the first PCa case was described by the surgeon J. Adams in 1853, great advances were achieved in the detection and treatment of PCa (1). Metastatic PCa was found to respond to androgen ablation therapy by C. Huggins (40s), the Gleason pathological grading system was developed (60s) and prostate specific antigen (PSA) was found to be elevated in the serum of PCa patients (80s). Although much progress has been made during the last 20 years, the molecular mechanisms underlying PCa development and progression are still not entirely understood.

In order to expand the understanding of the molecular events responsible for the initiation and progression of PCa, we used and combined genotyping, sequencing and gene expression data of several PCa samples. Our main objective was identifying novel genetic alterations that could provide mechanistic insight into these processes. Initially, we focused on the discovery of novel DNA and RNA fusions that would be frequently observed across our set of PCa patient samples. In this process, we also attempted to determine common gene mutations as a result of mismatch repair (MMR) deficiency in PCa.

Novel genetic alterations in PCa: – the era of rare genetic events

Like all cancers, prostate cancer is a genetic disease defined by genomic alterations ranging from point mutations to microsatellite variations and chromosomal alterations (translocations, insertions, duplications, deletions) (2). Already in the early days of Cellular Biology, the process leading to abnormal cell division and 'cancer' intrigued scientists. One of the greatest biologists of the twentieth-century, Theodor Boveri, proposed much ahead of its time, that chromosomal aberrations were the cause of cancer and predicted the existence of tumour suppressors and oncogenes (3, 4). His concept of genetic (chromosomal) instability has been further developed by several investigators in the past decades. In 2000, Hanahan and Weinberg proposed the first six hallmarks of cancer, consisting of features that most, if not all cancers require, to develop and progress (5). Later, following the broader progress in the cancer research field, the hallmarks were expanded to a total of ten (6).

In PCa, the discovery of recurrent gene fusion events contradicted the assumption at that time, that gene fusions were genetic alterations common in haematological malignancies and sarcomas. Gene fusions involving *TMPRSS2* and the *ETS* gene family of transcription factors were detected at high frequency in PCa (7). Following the identification of the *TMPRSS2-ERG* gene fusion, which is present in approximately 50% of all PCa cases (8), other less frequent gene fusions were identified involving members of the RAF family or other members of the ETS family (9). Therefore, we aimed at analysing both the genome and transcriptome of PCa samples in order to identify novel common gene fusions in PCa.

We identified two novel in-frame gene fusions involving members of the subfamily of membrane palmitoylated proteins (MPP), *MPP2* and *MPP5. MPP2* was found to be fused to *GPS2* (chapter 2). *MPP5* is fused to *FAM71D* and both genes are located in chromosome 14, adjacent to each other (chapter 3). Much like most gene fusions identified so far, with the exception of *TMPRSS2-ERG* itself, these particular gene fusions are rare. *MPP2* and *MPP5* are members of the MPP subfamily of the Membrane-Associated Guanylate Kinase (MAGUK) family of proteins to which the Disc large (Dlg) tumour suppressor gene also belongs (10). *GPS2* is a suppressor of G protein-activated MAPK signalling and is a stable component of the SMRT co-repressor complex (11). *FAM71D* is a largely unknown gene encoding a protein with unidentified protein domains. Whereas *GPS2* and *MPP2* have been, in some way, associated with cancer, nothing is known so far regarding *MPP5* and *FAM71D* (12, 13).

In contrast to most gene fusions identified in PCa, we did not find any evidence of androgen regulation of the upstream regulatory partner genes *GPS2* and *MPP5*. In any case, androgen regulation of gene fusions in PCa is not a requirement for functional relevance as shown for fusions involving *HNRPA2B1, DDX5* and *ESRP1* (fused to RAF1). *GPS2-MPP2* and the *MPP5-FAM71D* gene fusions were shown to increase proliferation and in the case of *GPS2-MPP2* also decrease apoptosis. As both evading apoptotic signals and increased proliferation capacity are acquired characteristics of tumour cells, *GPS2-MPP2* and also *MPP5-FAM71D* may provide a positive selective advantage towards malignant transformation.

The observation that several members of the MPP subfamily are affected by gene rearrangements supports a role of these proteins in PCa as has been shown for MPP3 in hepatocellular carcinoma (14). Since MPP members are shown to regulate both the trafficking and processing of nectins, important cell-cell adhesion molecules, they might play a role in PCa invasion (15). Nectins are expressed in various tumours and are currently being explored not only as potential targets for antibody therapy (16), but also prognostic markers like PVRL1 (Nectin-1) in PCa (17). So far, both MPP3 and MPP5 were shown to directly interact with nectins in an isoform-specific manner. This association between MPP members and nectins might be an important mechanism mediating PCa progression.

Interestingly, MPP proteins have also been shown to function as specific phospho-protein-binding modules (18). This translates into phosphorylation-dependent interactions between MPP and other proteins. Such proteins might include known tyrosine kinases such as protein kinase A (PKA), aPKC, PKB, and AKT that potentially regulate the assembly of large protein complexes by MAGUKs (18). Future research into the MPP subfamily of proteins and their role in PCa might identify novel signalling pathways or substantiate the importance of signalling pathways already known to be predominant in PCa.

Not only for PCa but also in several other cancer types, the use of next generation sequencing (NGS) technologies increased the number of somatic alterations detected. For the most part, this encompasses the identification of passenger events that do not contribute, as far as we know, to the biology of cancer. Determining which of the alterations detected are relevant, requires experimental validation and is currently a challenge due to the massive amounts of data generated by NGS. In addition, the majority of alterations identified by NGS are rare events with some specific to just one patient sample. This also represents a challenge for developing novel PCa biomarkers and novel effective therapies that can be used in a broad spectrum of patients. In the future, determining which pathways are commonly affected due to different mutations will contribute to solving the problem.

Gene mutations in prostate cancer – PRRT2 as a target of mismatch repair deficiency

The DNA mismatch repair system (MMR) is critical for preserving genomic stability through the recognition and repair of mismatch base-pairs that occur during DNA replication (19). MMR deficiency is a common characteristic of many tumour types since failure to repair the replication-associated errors results in a hypermutated phenotype with abnormally high mutation rates (20). This hypermutated phenotype has been described particularly for colorectal (CRC) and endometrial cancer (EC) (21, 22). Most of the hypermutated samples exhibited high microsatellite instability (MSI-H) caused by mutations in MMR genes or hypermethylation of *MLH1* (23).

In PCa, a hypermutated phenotype has also been observed (24). Subsequent exome sequencing of advanced PCa found this phenotype in 12% of the cases, all with loss of MMR and MSI-H (25). The MMR deficiency observed in PCa is caused essentially by mutations and complex structural rearrangements involving the *MSH2* and *MSH6* genes (25). So far, very few studies have focused on the prevalence of the hypermutated phenotype in PCa and the underlying mechanism. To further investigate the importance of MMR in PCa we have analysed whole genome sequencing data of the MMR-deficient PC346C cell line to determine the gene mutations resulting from the lack of MMR (chapter 4).

We identified *PRRT2* as homozygously affected in PC346C cells and frequently mutated in CRC and EC MSI-H patients. *PRRT2* is a poorly characterized protein that has been associated with several paroxysmal disorders (26). The mutation detected occurs in a 9C repeat in the second exon of *PRRT2* causing a frameshift that creates a premature stop codon. This 9C repeat is also known as the main mutation site for paroxysmal disorders (27). The low prevalence of paroxysmal disorders hampers a direct association between the *PRRT2* mutation and cancer. We determined that this *PRRT2* mutation generates a truncated protein that promotes cellular proliferation and migration. This gain of function might explain the high mutation frequency observed in CRC and EC patients.

The identification of *PRRT2* as being frequently mutated in MSI-H CRC and EC represents the first step in determining whether there are benefits from detecting this mutation. Although we have observed consistently and across different tumour types that the expression of *PRRT2* is lower in the tumour sample as compared to normal, the significance behind this fact remains undetermined. On one hand, such a differential expression pattern provides an ideal opportunity to explore *PRRT2* as a biomarker. On the other hand, we do not know whether having low expression of *PRRT2* is correlated to having the mutation in the 9C repeat or whether it associates with a poorer clinical outcome.

Although MSI-H has been associated with a slightly better prognosis in CRC, this area is still subject to great debate, especially since several study designs did not have sufficient power to infer such associations. Whether there is value in determining the *PRRT2* status independently of MSI-H is unclear. It is known for CRC that MMR deficiency can cause mutations in important tumour suppressor genes like *TGFβR2* (28). Depending on whether these MMR secondary gene mutations can be therapeutically targetable, it might be interesting to determine the mutation status of such genes (29).

The genomic landscape of prostate cancer – is there a role for chromothripsis?

As genome sequencing technologies become readily available and affordable, also the number of novel somatic mutations identified in cancer grows. However, identifying novel gene mutations is not the only end result of such informative data. Another advantage of large-scale high resolution genome studies is the possibility to identify mechanisms driving the mutation signatures seen in cancer.

One major leap in the field of molecular biology made possible by NGS technologies was the discovery of 'kataegis' (mutation storms) and 'chromothripsis' (chromosomal crisis) (30, 31). Kataegis represents a pattern of localized hypermutation. Chromothripsis is the process by which a genome shatters into hundreds of fragments from one or several chromosomes in a single cellular event and further recombines these fragments in a highly erroneous order. Both kataegis and chromothripsis represent new oncogenic mechanisms that challenge the classic model of a gradual and stepwise carcinogenesis process.

The sequencing of the VCaP cell line showed that this rather peculiar phenomenon of chromothripsis is also present in PCa. The premise that, by such a mechanism, multiple gene amplifications/deletions and rearrangements can activate oncogenes and/or inactivate tumor suppressors, prompted us to identify gene fusions caused by this phenomenon. We detected a very limited number of same-direction gene fusions (15 events) in comparison to the total number of 573 rearrangements and only two were predicted to be in-frame. Clearly, the formation of gene fusions as a consequence of chromothripsis is not preferential over rearrangements occurring outside genes, neither is there a positive selection for in-frame transcripts.

Chromosomal crises have been observed in 2 to 3% of the most common cancer types and are particularly frequent in bone cancer (35). The definition of chromothripsis presents a challenge as to how cells, undergoing such dramatic genetic catastrophes, can survive. So far, it has been linked to poor prognosis in acute myeloid leukaemia, neuroblastoma and multiple myeloma (36). If more reports are to substantiate this link between chromothripsis and prognosis, there might be potential for its use in aiding prognosis and possibly patient stratification.

A collection of 22,347 cancer genome screens has been used to catalogue the occurrence of chromothripsis-like patterns (CTLP) (32). From this study, there seems to be a suggestive poorer overall survival for cancer patients with CTLP as compared to non-CTLP. In a large study of 8000 cancer genomes, there was evidence that from all the epithelial tumors analyzed, PCa shows the highest frequency of CTLP (21 affected samples out of a total number of 372) (33). A later study by Baca et al. suggests 'chromoplexy' instead of chromothripsis to be much more common in PCa (34). Chromoplexy is essentially a smaller version of chromothripsis (less rearrangements) in sequential order as opposed to a single event. Clearly, further research in this field is required to clarify the definitions of both novel mechanisms and the mechanistic differences between chromoplexy and chromothripsis.

In the near future, the number of sequenced cancer genomes will certainly multiply. This will be especially important to accurately define how frequent all these new phenomena such as kataegis, chromothripsis and chromoplexy really are. Regarding PCa, it is essential to determine whether chromothripsis is correlated to a more initial or later stage of the disease. If indeed as recently suggested, chromothripsis is somehow associated with chronic prostatitis (35) it might be interesting to study this further in the context of PCa development. Also, with the potential role of the AR in promoting gene rearrangements, there might be an underlying link between AR and chromothripsis in PCa (36).

Advances in DNA sequencing technologies – what does the future hold?

As described above, over the past years an extraordinary development of molecular profiling techniques was achieved, up to the most recent boost of next generation sequencing. Initially, the study of the genetic basis of Mendelian disorders was carried out through linkage mapping and cytogenetic analysis. But these techniques have a low resolution and therefore, a fine description of genetic alterations is virtually impossible. With the development of DNA microarray technology, genetic alterations could be observed at a relatively higher resolution (15-35 kb) through array comparative genomic hybridization (aCGH). Progressively, the use of aCGH was replaced by single nucleotide polymorphisms-based platforms (SNPs) that allowed both the estimation of genome-wide variations in copy number changes and allelic contents at an average resolution of 3 kb (37, 38). This increased resolution allowed uncovering whether DNA breaks occurred inside genes.

On the RNA level, microarrays revolutionized gene expression analysis from a single gene to all genes in a single experiment. This technology allowed for the monitoring of mRNA abundance which in turn provides complete information about the cell's 'transcriptome'. Due to the dynamic characteristics of the cell's transcriptome, microarrays are very powerful in creating multi-gene patterns of expression for different cancer types, contexts, treatments, etc. One such example was the distinction between acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL) solely based on microarray transcript data (39). Also regarding the RNA level, the 'next-generation' revolution has arrived. Currently, NGS technologies are used to sequence cDNA enabling the discovery of novel genes and detection of alternative splicing and gene fusions (40).

Since the publication of the first draft of the human genome (41, 42), tremendous advances were achieved regarding sequencing technologies and by now, NGS has mostly replaced the use of DNA microarrays. In the past few years, the sequencing of the first cancer genome in 2008 by Ley TJ *et al*. (43) was expanded to more than 800 fully sequenced genomes of at least 25 different cancer types (44). Up to 2014, the total estimated number of human genomes sequenced was approximately 228,000 according to Francis de Souza, the president of Illumina Inc.

Despite the growing number of novel somatic mutations identified and some proven driver mutations, few have been translated into new treatment discovery and development. One of the most successful examples is the identification of frequent *EZH2* mutations in lymphoma with promising treatment results in pre-clinical studies (33). Nonetheless, one of the most important contributions of NGS to the cancer field has been the challenging of old concepts. The high resolution of NGS allowed the characterization of clonal evolution in the tumorigenesis process (45).

In PCa, it was shown earlier this year, that morphologically 'normal' cells, already contain mutations that are present in distant tumour tissue (46). This is in agreement with the idea of carcinogenesis as a multistep process with separate clonal cell expansions as described by Vogelstein for colorectal cancer (47). Also, a recent study by Gundem *et al.* described the evolution process of metastatic PCa and showed that overall, there is a common shared heritage between both primary tumour and metastasis and that metastasis occurs as a spread of multiple tumour clones at the same time across the whole body (48).

We can expect that in the next years more and more NGS studies will reveal how tumorigenesis actually develops in a temporal and spatial manner. This will shed light into the true complexity that lies within rather straightforward phenotypes and/or genetic alterations. Especially considering heterogeneous tumours like PCa where different primary tumour foci harbour different genetic alterations, it is important to understand what changes confer an aggressive phenotype and require further attention concerning the clinical outcome.

In the future, sequencing costs will probably decrease further from the current $3,000 per genome. The drop will likely not be as sharp as it has been until now but it will surely make the use of DNA studies more affordable for use in the clinic as general practice. DNA sequencing might become routine in medicine with the approval of diagnostic tests that combine sequencing of the most relevant cancer genes. So far, the major bottleneck in widening NGS technologies to common practice is the storing and analysis of such massive amounts of data. Deciphering what really matters from what does not, depends not only on which question is being asked but also on the extend of our knowledge in the subject.

Prostate cancer treatment and diagnosis – are we getting close to personalized medicine?

The concept of precision and personalized medicine has always been an intrinsic part of the regular physician's practice. Medical doctors have always tried to find the best match between patient and treatment, much like the search for suitable blood, bone marrow and organ donors. It is important to consider not only the environment, health and behaviour of the patient, but also his genomic background, hopefully achieving higher accuracy in personalized treatments. In the prostate cancer field, the discovery of androgen dependence in 1941 by Huggins and colleagues, represented a major milestone in developing treatment for this disease. Throughout the years, the AR remained the backbone in treatment of metastatic PCa but there is a clear need for more therapy targets and treatment options.

The challenge in developing novel therapies has mostly to do with the druggability of the genetic alteration. So far, the majority of successful cancer targeted therapies are small-molecule kinase inhibitors. In PCa, drugs targeting relevant signalling pathways other than the AR such as PI3K/Akt, HDAC and Src have failed in showing a significant improvement in disease outcome (49). Surely, continuing the search for alternative signaling pathways that drive prostate cancer progression will provide an alternative landscape for therapeutic treatment. Also, using different treatment combinations or known treatments but in a different order might improve the disease outcome.

In PCa, most of the biomarker research is based on the diagnosis and prognosis of the disease. This is based on a great need for biomarkers that help distinguish BPH from PCa, indolent from aggressive disease and identify metastatic PCa (50). Due to the heterogeneity of PCa and the presence of different genetic abnormalities between different foci in the same patient, it will remain hard to determine the true stage of the disease

Until now, several gene fusions have been observed in PCa patients and, although these are mostly patient-specific, some do involve important cancer-related genes like *MAPK1, BRAF* and *RB1* (51). The identification of these novel genetic alterations can be useful in detection, monitoring or treatment of PCa patients. The presence of cancer-specific markers like *TMPRSS2-ERG* gives a clear answer to whether a patient has PCa.

But even if a novel marker can be validated, it will be its proven added-value that will determine, for most part, whether it will be implemented in the clinic or not. We have come a long way regarding the discovery of novel candidate markers and now the focus should be on creating better validation and clinical implementation approaches.

The future of PCa patient diagnosis and treatment will most likely rely on a 'personalized oncology – precision medicine' approach where each patient's tumour DNA will be screened to determine the individual treatment strategy. This inter-vention method will include the use of early diagnostic biomarkers that will, in the first instance, serve as a screening tool to decide which patients should undergo tissue sampling and further DNA sequencing (10). Is there cancer? Should it be treated? And if so, which treatment, order and/or combination should be implemented? The generation of large scale sequencing maps of PCa patients provides a first attempt to answer these questions.

References

1. Denmeade, S.R. and J.T. Isaacs, A history of prostate cancer treatment. Nat Rev Cancer, 2002. 2(5): p. 389-96.
2. Boyd, L.K., X. Mao, and Y.J. Lu, The complexity of prostate cancer: genomic alterations and heterogeneity. Nat Rev Urol, 2012. 9(11): p. 652-64.
3. Balmain, A., Cancer genetics: from Boveri and Mendel to microarrays. Nature Reviews Cancer, 2001. 1(1): p. 77-82.
4. Wunderlich, V., "He Corrects my View and Develops it Further." Comments by David von Hansemann on the Monograph by Theodor Boveri Concerning the Origin of Malignant Tumors (1914). Berichte Zur Wissenschaftsgeschichte, 2011. 34(3): p. 263-283.
5. Hanahan, D. and R.A. Weinberg, The hallmarks of cancer. Cell, 2000. 100(1): p. 57-70.
6. Hanahan, D. and R.A. Weinberg, Hallmarks of cancer: the next generation. Cell, 2011. 144(5): p. 646-74.
7. Tandefelt, D.G., et al., ETS fusion genes in prostate cancer. Endocrine-Related Cancer, 2014. 21(3): p. R143-R152.
8. Tomlins, S.A., et al., Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science, 2005. 310(5748): p. 644-648.
9. Palanisamy, N., et al., Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nature Medicine, 2010. 16(7): p. 793-U94.
10. de Mendoza, A., H. Suga, and I. Ruiz-Trillo, Evolution of the MAGUK protein gene family in premetazoan lineages. BMC Evol Biol, 2010. 10: p. 93.
11. Oberoi, J., et al., Structural basis for the assembly of the SMRT/NCoR core transcriptional repression machinery. Nat Struct Mol Biol, 2011. 18(2): p. 177-84.
12. Wong, M.M., C. Guo, and J. Zhang, Nuclear receptor corepressor complexes in cancer: mechanism, function and regulation. Am J Clin Exp Urol, 2014. 2(3): p. 169-87.
13. Baumgartner, M., et al., The PDZ protein MPP2 interacts with c-Src in epithelial cells. Exp Cell Res, 2009. 315(17): p. 2888-98.
14. Ma, H., et al., Membrane palmitoylated protein 3 promotes hepatocellular carcinoma cell migration and invasion via up-regulating matrix metalloproteinase 1. Cancer Lett, 2014. 344(1): p. 74-81.
15. Dudak, A., et al., Membrane palmitoylated proteins regulate trafficking and processing of nectins. Eur J Cell Biol, 2011. 90(5): p. 365-75.
16. Oshima, T., et al., Nectin-2 is a potential target for antibody therapy of breast and ovarian cancers. Mol Cancer, 2013. 12: p. 60.
17. Kalin, M., et al., Novel prognostic markers in the serum of patients with castration-resistant prostate cancer derived from quantitative analysis of the pten conditional knockout mouse proteome. Eur Urol, 2011. 60(6): p. 1235-43.
18. Zhu, J., et al., Guanylate kinase domains of the MAGUK family scaffold proteins as specific phospho-protein-binding modules. EMBO J, 2011. 30(24): p. 4986-97.
19. Li, G.M., Mechanisms and functions of DNA mismatch repair. Cell Res, 2008. 18(1): p. 85-98.
20. Bak, S.T., D. Sakellariou, and J. Pena-Diaz, The dual nature of mismatch repair as antimutator and mutator: for better or for worse. Frontiers in Genetics, 2014. 5: p. 287.
21. Cancer Genome Atlas, N., Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012. 487(7407): p. 330-7.

22. Cancer Genome Atlas Research, N., et al., Integrated genomic characterization of endometrial carcinoma. Nature, 2013. 497(7447): p. 67-73.
23. Roberts, S.A. and D.A. Gordenin, Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer, 2014. 14(12): p. 786-800.
24. Kumar, A., et al., Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci U S A, 2011. 108(41): p. 17087-92.
25. Pritchard, C.C., et al., Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. Nat Commun, 2014. 5: p. 4988.
26. Meneret, A., et al., PRRT2 mutations and paroxysmal disorders. Eur J Neurol, 2013. 20(6): p. 872-8.
27. Chen, W.J., et al., Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. Nat Genet, 2011. 43(12): p. 1252-5.
28. Kloor, M., [Pathogenesis of microsatellite-unstable colorectal cancer. Evaluation of new diagnostic and therapeutic options]. Pathologe, 2013. 34 Suppl 2: p. 277-81.
29. Guillotin, D. and S.A. Martin, Exploiting DNA mismatch repair deficiency as a therapeutic strategy. Experimental Cell Research, 2014. 329(1): p. 110-115.
30. Stephens, P.J., et al., Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell, 2011. 144(1): p. 27-40.
31. Nik-Zainal, S., et al., Mutational processes molding the genomes of 21 breast cancers. Cell, 2012. 149(5): p. 979-93.
32. Cai, H., et al., Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens. BMC Genomics, 2014. 15: p. 82.
33. Kim, T.M., et al., Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res, 2013. 23(2): p. 217-27.
34. Baca, S.C., et al., Punctuated evolution of prostate cancer genomes. Cell, 2013. 153(3): p. 666-77.
35. Lapuk, A.V., et al., From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. J Pathol, 2012. 227(3): p. 286-97.
36. Haffner, M.C., et al., Androgen-induced TOP2B mediated double strand breaks and prostate cancer gene rearrangements. Nature genetics, 2010. 42(8): p. 668-675.
37. Carter, N.P., Methods and strategies for analyzing copy number variation using DNA microarrays. Nature Genetics, 2007. 39: p. S16-S21.
38. Ionita-Laza, I., et al., Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics, 2009. 93(1): p. 22-6.
39. Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999. 286(5439): p. 531-7.
40. Guo, Y., et al., Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. PLoS ONE, 2013. 8(8): p. e71462.
41. Venter, J.C., et al., The sequence of the human genome. Science, 2001. 291(5507): p. 1304-+.
42. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.
43. Ley, T.J., et al., DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature, 2008. 456(7218): p. 66-72.
44. Mwenifumbo, J.C. and M.A. Marra, Cancer genome-sequencing study design. Nat Rev Genet, 2013. 14(5): p. 321-32.

45. Shah, S.P., et al., Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature, 2009. 461(7265): p. 809-U67.
46. Cooper, C.S., et al., Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat Genet, 2015.
47. Fearon, E.R. and B. Vogelstein, A genetic model for colorectal tumorigenesis. Cell, 1990. 61(5): p. 759-767.
48. Gundem, G., et al., The evolutionary history of lethal metastatic prostate cancer. Nature, 2015. advance online publication.
49. Chung, P.H., et al., Emerging drugs for prostate cancer. Expert Opin Emerg Drugs, 2013. 18(4): p. 533-50.
50. Prensner, J.R., et al., Beyond PSA: the next generation of prostate cancer biomarkers. Sci Transl Med, 2012. 4(127): p. 127rv3.
51. Barbieri, C.E., et al., The mutational landscape of prostate cancer. Eur Urol, 2013. 64(4): p. 567-76.

Abbreviations

AATK	apoptosis-associated tyrosine kinase
aCGH	array comparative genomic hybridization
AR	androgen receptor
ARHGEF3	rho guanine nucleotide exchange factor (GEF) 3
BRAF	B-Raf proto-oncogene, serine/threonine kinase
C8ORF38	chromosome 8 open reading frame 83
CEP164	centrosomal protein 164kDa
CHD1	chromodomain helicase DNA binding protein 1
CNOT1	CCR4-NOT transcription complex, sub 1
CNVs	copy number variations
COSMIC	catalogue of somatic mutations in cancer
CRC	colorectal cancer
CRPC	castration-resistant prostate cancer
CTLP	chromothripsis-like patterns
DAB2IP	DAB2 interacting protein
DLG	discs large homolog
DNR	dinucleotide repeat
DuCaP	dura mater-Cancer of the Prostate (cell line)
EC	endometrial cancer
ERG	v-ets erythroblastosis virus E26 oncogene homolog (avian)
ETS	E26 transforming sequence
FAM71D	family with sequence similarity 71, member D
FISH	fluorescent in situ hybridization
GPS2	G protein pathway suppressor 2
GWAS	genome-wide association studies
HGPIN	high grade prostatic intraepithelial neoplasia
LNCaP	lymph node-Cancer of the Prostate (cell line)
MAGUK	membrane-associated guanylate kinase
MAPK	mitogen-activated protein kinase
MAST4	microtubule associated serine/threonine kinase family member 4
MMR	mismatch repair
MNR	mononucleotide repeat
MPP	membrane palmitoylated protein
MPP2	membrane palmitoylated protein 2
MPP5	membrane palmitoylated protein 5
MSI	microsatellite instability
NCOA2	nuclear receptor coactivator 2
N-COR	nuclear receptor corepressor
NGS	next generation sequencing
NRXN3	neurexin 3
PCa	prostate cancer
PCA3	prostate cancer gene 3
PI3K	phosphoinositol-3-kinase

PKD	paroxysmal kinesigenic dyskinesia
PRRT2	proline-rich transmembrane protein 2
PSA	prostate specific antigen
PTEN	phosphatase and tensin homolog
SMRT	silencing mediator for retinoid and thyroid hormone receptors
SNPs	single nucleotide polymorphisms
SVs	structural variations
TMPRSS2	transmembrane protease, serine 2
TNR	trinucleotide repeat
TP53	tumor protein P53
TSHR	thyroid stimulating hormone receptor
TSHZ2	teashirt zinc finger homeobox 2
TURP	transurethral resections of the prostate
VCaP	vertebral-Cancer of the Prostate (cell line)

Summary

Prostate cancer (PCa) is a common disease of the western society. The risk of developing PCa increases with age and most men in their 80s will have developed PCa, even if the disease is not clinically relevant. Contrary to indolent PCa, aggressive disease has a poor survival outcome. Better treatment options and better patient stratification are urgent needs in the PCa field. In this thesis we have focused on the identification of novel genetic alterations in PCa. In the long term, identification of novel abnormalities can aid identifying specific subsets of patients or novel targetable pathways for the treatment of this disease.

In chapter 1, a general introduction is provided, with special emphasis on the major signaling pathways affected in prostate cancer and the developments in technology together with emerging markers.

To identify novel genetic alterations in PCa we used, in chapter 2, data from DNA and RNA microarrays performed in a large dataset of PCa samples. The initial aim was to detect novel DNA and RNA gene fusions. DNA microarray data was analysed with the Nexus software to obtain a list of all DNA break locations in our samples to determine which DNA breaks were located inside genes. Using the RNA expression data we determined which genes had both DNA breaks and abnormal expression. The top 3 candidates were used to identify possible fusion partners through 5'RACE PCR. Following this approach we found two novel fusions in PCa, *GPS2-MPP2* in the LNCaP cell line and *TSHR-NRXN3* in the G295 PCa clinical sample. *GPS2-MPP2* was further investigated regarding its functional consequences and shown to promote proliferation and suppress apoptosis in LNCaP cells. In conclusion, combining DNA and RNA microarray data is a powerful approach to determine novel gene fusions but additional validation of the findings is fastidious.

The boost of next generation sequencing (NGS) techniques permitted the use of this technology to evaluate, in chapter 3, if through a whole genome sequencing approach, the number of gene fusions identified in prostate cancer could be increased. Overall, most of the 7735 breaks identified in two genomic DNA samples, cell line PC346C and clinical sample G089, turned out to be polymorphisms present also in a control dataset of 46 normal samples. The total number of cancer-related rearrangements identified was 674 and 387 in PC346C and G089, respectively. From these, only 14 and 33 occurred between two different gene locations and only 20 were predicted to generate feasible gene fusions. Through PCR and Sanger sequencing 17 out of the 20 could be validated. Most of the genes involved in these fusions had not previously been associated with cancer and only a small subset of these was also found to have additional breaks in other PCa samples that were genotyped in chapter 2. The in-frame *MPP5-FAM71D* gene fusion resulted in an upregulation of *FAM71D* expression, which promoted the proliferation of PC346C cells. Despite the finding that the *MPP5-FAM71D* gene fusion seemed to be specific for PC346C, 10 other PCa samples showed also *FAM71D* upregulation. In this way, whole genome sequencing does provide a more comprehensive catalog of

structural variations. However, the identification of frequent and functionally relevant fusions requires large datasets of patient samples and the whole genome sequencing of such large numbers remains costly.

As PC346C is a mismatch repair (MMR) deficient cell line and whole genome sequencing data was available, we examined whether there was an association between MMR deficiency and mutations in nucleotide repeat sequences in PC346C. We observed a bias towards mutations in mononucleotide repeats (MNR) in the PC346C which was not present in the mismatch repair proficient PCa sample G089. To further evaluate whether this bias had functional relevance, we generated a list of genes mutated in MNR in the PC346C sample. The top candidates, with no previous association with mismatch repair and microsatellite instability (MSI), were validated in a large panel of MSI stable and unstable cell lines, namely prostate, colorectal, endometrial and ovarian cancer cell lines. A mutation in a 9C repeat of the *PRRT2* gene was found to be frequent in several of the cancer cell lines tested as well as in MSI unstable colorectal and endometrial patients. Through overexpression of both the wildtype *PRRT2* and the mutated *PRRT2* we observed an increase in the proliferation of cells in the presence of mutated *PRRT2*. Also, the migratory capacity of cells was greatly increased. The overexpression of the wildtype *PRRT2*, on the contrary, caused a mild decrease in proliferation with no significant effect on migration. *PRRT2* expression is, in general, lower in cancer as compared to normal, not only in PCa but also in lung and gastric adenocarcinomas. In the end, we consider mutated *PRRT2* to be a novel dominant oncogene, and wt*PRRT2* a tumor suppressor candidate.

The latest developments in cancer research had identified, based on NGS data, a phenomenon called chromothripsis. The identification of such phenomenon, which consists of tens to hundreds of rearrangements during a single catastrophic event, was made possible through high resolution whole genome NGS. Among our set of whole genome sequenced PCa samples, we observed the occurrence of a massive number of rearrangements in chromosome 5 of the VCaP PCa cell line. Following closer examination, we identified 573 DNA breakpoints in the long arm of chromosome 5, of which only 4 involved additional chromosomes. Of these 573 DNA breaks, 43 involved different genes on both sides of the break and only 18 of these events could generate gene fusions. By qPCR we were able to confirm at the DNA level, 15 gene fusions but we could only detect 5 gene fusion transcripts. In the end, the only in frame fusion transcript identified that resulted from the chromothripic event was *NDUFAF2-MAST4*. Whether *NDUFAF2-MAST4* is of functional relevance in the initiation or progression of PCa was not tested but our data indicated that chromothripsis does not preferentially result in novel gene fusions.

To conclude, through the use of several high throughput technologies, both at the DNA and RNA level, we were able to identify novel genetic alterations in PCa. None of the genetic alterations characterized in this thesis occured frequently in clinical PCa samples. The individual character of these alterations does not necessarily signify they are ineffectual; instead they might reveal the importance of common pathways, targeted by the neoplastic process at different levels.

Samenvatting

Prostaat kanker (PKa) is een frequent voorkomende ziekte bij mannen in de Westerse wereld. Het risico om PKa te ontwikkelen wordt groter naarmate men ouder wordt en na hun 80e levensjaar zullen de meeste mannen PKa hebben, ook al is dit meestal niet klinisch relevant. In tegenstelling tot indolente PKa, geeft agressieve PKa slechts een kleine kans op overleving. Betere behandelingen en indeling van patiënten in groepen met verschillen in prognose zijn noodzakelijke ontwikkelingen in het PKa veld. In dit proefschrift hebben wij ons gefocust op de identificatie van nieuwe genetische veranderingen in PKa. Op de lange termijn kan de ontdekking van nieuwe afwijkingen helpen bij het identificeren van nieuwe doelwitten voor therapeutische behandeling en bij de selectie van patiënten die gebaat zijn bij behandeling van deze ziekte.

Het eerste hoofdstuk beschrijft een algemene introductie met specifieke nadruk op netwerken van signaaloverdracht die veranderd zijn in PKa en van de technologische ontwikkelingen bij het onderzoek naar nieuwe markers.

Om nieuwe genetische afwijkingen in PKa te identificeren, hebben wij in hoofdstuk 2 gegevens gebruikt van DNA en RNA microarrays van een groot aantal PKa monsters. Het aanvankelijke doel was om nieuwe DNA en RNA fusiegenen te detecteren. De DNA microarray gegevens werden geanalyseerd met Nexus software om een lijst te verkrijgen van alle locaties van DNA breuken in onze monsters en om te bepalen welke DNA breuken zich in genen bevonden. Gebruik makend van de RNA expressie data, hebben wij omschreven welke genen zowel breuken in het DNA als abnormale expressie vertoonden. De top 3 kandidaten werden gebruikt om potentiele fusiepartners te ontdekken met behulp van een 5'RACE PCR. Met deze aanpak hebben wij twee nieuwe fusie genen ontdekt in PKa, *GPS2-MPP2* in de LNCaP cellijn en *TSHR-NRXN3* in het klinische PKa monster G295. *GPS2-MPP2* werd verder onderzocht op functionele consequenties. Het fusiegen bleek proliferatie te stimuleren en apoptose te onderdrukken in LNCaP cellen. Samenvattend, het combineren van DNA en RNA microarray data is een krachtige techniek om nieuwe fusie genen te ontdekken, maar de validatie van deze bevindingen komt erg nauw.

De sterke ontwikkeling van next generation sequencing (NGS) technieken heeft ertoe geleid, dat wij NGS konden toepassen bij het evalueren van een lijst van fusiegenen en bij het uitbreiden hiervan (hoofdstuk 3). Globaal gezien waren de meeste van de 7735 geïdentificeerde breuken in twee onderzochte PKa samples polymorfismen, die tevens aanwezig waren in een controle dataset van 46 normale monsters. Het totale aantal aan PKa gerelateerde breuken dat ontdekt werd was respectief 674 en 387 in PC346C en G089. Hiervan waren slechts 14 en 33 aanwezig tussen twee verschillende gen locaties en van slechts 20 kon worden voorspeld dat ze genfusies konden betreffen. Door middel van PCR en Sanger sequencing konden 17 van de 20 bevestigd worden. De meeste genen betrokken bij deze fusies waren nog niet eerder geassocieerd met kanker en slechts een gedeelte ervan had breuken in andere PKa monsters, gegenotypeerd in hoofdstuk 2. De in-frame *MPP5-FAM71D* genfusie veroorzaakte een verhoogde expressie van *FAM71D*,

dat leidde tot een verhoogde proliferatie van PC346C cellen. Ondanks het feit dat de *MPP5-FAM71D* fusie specifiek leek voor PC346C, vertoonden 10 andere PKa monsters ook verhoogde expressie van *FAM71D*. Op deze wijze biedt whole genome sequencing een uitgebreide catalogus van structurele variaties. De identificatie van frequente en functioneel relevante fusies, verlangt echter grotere datasets van patiënt monsters en whole genome sequencing van zulke grote aantallen, blijft prijzig.

Aangezien PC346C een mismatch repair (MMR) deficiënte cellijn is en whole genome sequencing beschikbaar was, hebben wij de associatie tussen MMR deficiëntie en mutaties in nucleotide repeat sequenties in PC346C onderzocht. Wij observeerden een voorkeur voor mutaties in mononucleotide repeats (MNR) in PC346C, die niet aanwezig was in het mismatch repair competente PKa monster G089. Om verder te onderzoeken of deze voorkeur functioneel relevant was, hebben wij een lijst gemaakt van genen gemuteerd in MNR in PC346. De top kandidaten, die geen eerdere associatie hadden met mismatch repair en microsatelliet instabiliteit (MSI), werden bevestigd in een uitgebreide selectie van MSI stabiele en instabiele cellijnen, namelijk: prostaat, colorectaal, endometrium en eierstok kanker cellijnen. Een mutatie in een 9C repeat van het *PRRT2* gen bleek frequent voor te komen in meerdere van de geteste kanker cellijnen en in de MSI instabiele colorectale en endometriale patiëntensamples. Door middel van overexpressie van zowel wildtype *PRRT2* als gemuteerde *PRRT2*, hebben we een verhoogde proliferatie geconstateerd in de aanwezigheid van het gemuteerde *PRRT2* gen. De capaciteit tot migreren van cellen was tevens sterk verhoogd. Integenstelling hiermee veroorzaakte overexpressie van wildtype *PRRT2* een beperkte vertraging van proliferatie en had geen significant effect op migratie. *PRRT2* expressie was over het algemeen lager in kanker dan in normaal weefsel, niet alleen PKa, maar ook in long- en maagcarcinomen. Wij beschouwen gemuteerd *PRRT2* als een nieuw, dominant oncogen en wt*PRRT2* als een mogelijke tumor suppressor.

De laatste ontwikkelingen in kankeronderzoek hebben, gebaseerd op NGS data, een fenomeen ontdekt genaamd chromothripsis. De identificatie van een dergelijk fenomeen, dat tien tot honderden genomische herverdelingen omvat ontstaan bij een enkele gebeurtenis, is mogelijk gemaakt door hoge resolutie whole genome NGS. In onze set van whole genome sequenced PKa monsters hebben wij massale herverdelingen ontdekt in chromosoom 5 van de VCaP PKa cel lijn. Nadere inspectie liet zien dat 573 DNA breukpunten in de lange arm van chromosoom 5 aanwezig waren, waarvan slechts bij 4 van deze breuken andere chromosomen waren betrokken. Van deze 573 DNA breuken waren er bij 43 verschillende genen betrokken aan beide zijden van de breuk en slechts 18 konden genfusies genereren. Met een kwantitatieve PCR konden we 15 genfusies bevestigen op het DNA niveau, maar we konden slechts 5 transcripten van fusiegenen detecteren. Uiteindelijk konden we het in frame fusie transcript *NDUFAF2-MAST4* als enige identificeren, als resultaat van de chromothripsis gebeurtenis. Wij hebben niet getest of *NDUFAF2-MAST4* functioneel relevant kan zijn bij het ontstaan of de progressieve groei van PKa, maar onze data suggereren dat chromothripsis niet noodzakelijk resulteert in genfusies.

Samenvattend: met behulp van verschillende high throughput technieken op zowel DNA als RNA niveau, konden wij nieuwe genetische veranderingen identificeren in PKa. Geen enkele genetisch verandering in deze thesis, was veelvoorkomend in klinische PKa monsters. Het individuele karakter van deze veranderingen betekent niet zozeer dat deze irrelevant zijn, ze kunnen mogelijk cruciale componenten zijn van de vele complexe netwerken van signaaloverdracht betrokken bij neoplastische processen.

Curriculum vitae

Inês Teles Siefers Alves was born on the 27th of December 1985 in Lisbon, Portugal. Also in Lisbon, she followed primary and secondary education. In the first year of high school she changed from Sciences to the Language and Humanities course. After finishing the first year she decided to go back to the Sciences course and finished high school with distinction. After graduation in 2004 she started her education in Biology in the Faculty of Sciences, Lisbon University. She obtained her Bachelor degree in 2007 and started the Human Molecular Biology Master program. During her University education she was granted two honor scholarships. In late 2009, she began her research internship in the laboratory of Prof. Dr. João Ferreira at the Molecular Medicine Institute in Lisbon. Here, she studied the role of BRCA1 and other chromatin structure regulators in the DNA damage response to topoisomerase-II inhibitors. In September 2009, she obtained her Master of Science diploma with excellence and decided to apply to the Marie Curie Initial Training Network fellowship in the group of Prof. Dr. Jan Trapman and Prof. Dr. Guido Jenster. Following acceptance she moved to Rotterdam and conducted her PhD research starting from November 2009 to May 2014. From May 2014 until now, she conducts her research as a post-doc in the group of Prof. Dr. Marcel van Vugt studying the genetic and proteomic interactions of PICH, a DNA helicase.

List of Publications

1) Teles Alves I, Hiltemann S, Hartjes T, van der Spek P, Stubbs A, Trapman J, Jenster G, Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. Human Genetics, 2013; 132:709–713

2) Teles Alves I, Hartjes T, McClellan EA, Temanni MR, Janssen B, van Workum W, van der Spek P, Stubbs A, de Klein A, Eussen B, Trapman J, Jenster G, Next generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. Oncogene, 2015; 34(5):568-77

3) Hiltemann S, McClellan EA, van Nijnatten J, Horsman S, Palli I, Teles Alves I, Hartjes T, Trapman J, van der Spek P, Jenster G, Stubbs A., iFUSE: integrated fusion gene explorer. Bioinformatics, 2013; 29(13):1700-1

PhD portfolio

Name PhD student: Inês Teles Siefers Alves **Erasmus MC Department:** Urology and Pathology **Research School:** Erasmus MC Graduate School	**PhD period:** 5 years **Promotor(s):** Prof. Jan Trapman and Prof. Guido Jenster **Supervisor:** Prof. Jan Trapman and Prof. Guido Jenster	
1. PhD training	**Year**	**Workload (Hours/ECTS)**
General courses		
- Time and project management training	2010	0.3
- Communications training workshop	2011	0.5
- Leadership and Entrepeneurship training	2011	0.6
- Job applications skills workshop	2012	0.4
- Grant application writing workshop	2012	0.2
Specific courses, seminars and workshops		
- Data analysis with Partek	2010	1.3
- Basic course on R	2010	1.4
- Nexus training course	2010	1
- Ensembl workshop	2010	0.6
- Photoshop CS3 workshop	2010	0.3
- Course on Biomedical reseach techniques	2010	1.5
- Basic and translational Oncology	2010	1.8
- Research management for PhD students	2010	1
Presentations		
- Chromothripsis in prostate cancer: what to make of it?" at the 2nd Daniel den Hoed day, Rotterdam, The Netherlands	2013	1
National and international conferences		
- 19th Meeting of the EAU Section on Urological Research (ESUR), Vilnius, Lithuania (Poster and Presentation)	2010	1
- 9th World Congress on Urological Research (WBUR), Innsbruck, Austria (Poster and Presentation)	2011	1
- 20th meeting of the EAU of Urological Research, Strasbourg, France (Poster and Presentation)	2012	1
- Prostate Action Forum, Rotterdam, The Netherlands	2012	1
- 21st Meeting of the EAU Section of Urological Research in collaboration with the 65th Congress of the DGU, Dresden, Germany (Poster and Presentation)	2013	1
- 15th Molecular Medicine day (Poster)	2011	1
- 16th Molecular Medicine day (Poster)	2012	1
- 17th Molecular Medicine day (Poster)	2013	1
2. Teaching	**Year**	**Workload (Hours/ECTS)**
Supervising practicals and excursions, Tutoring		
- Supervision of a Bachelor student for 6 months	2011	4
Supervising Masters theses		
- Supervision of a Master student for 9 months	2012	6

Acknowledgements

A tese de doutoramento é um caminho sinuoso e àrduo. Se não fosse a filosofía de Fernando Pessoa –"Pedras no caminho? Guardo todas, um dia vou construir um castelo…" – talvez não estivesse aqui este livro. Mas, como disse Friedrich Nietzsche e bem, – "Was ihn nicht umbringt, macht ihn stärker." E em consequência desses grandes senhores e grandes sabedorias, eu guardei todas as pedras que encontrei...e devo dizer que foram muitas...construi esta tese com todo o carinho que ainda consegui espremer do pouco que sobrou de mim depois desta jornada...e espero agora ser mais forte do que era antes – dado que ainda penso e se "Cogito, ergo sum" (René Descartes).

Este caminho foi àrduo não só para mim como também para todos os que me acompanharam e não seria senão justo deixar aqui, nestas ùltimas paginas, as merecidas palavras de gratidão a:

...Todos os pacientes e médicos que juntos tornaram possível o meu trabalho durante estes ùltimos anos, através do seu consentimento e colheita de material para investigação. O meu muito obrigado...To all the patients and doctors that have contributed to this research a sincere thank you.

To the thesis committee, thank you so much for carefully reading my thesis and giving much appreciated input. The end product is definitely better thanks to your assistance.

To everyone from the departments of Pathology and Urology. I will not mention each single person individually as I believe that everyone who has helped me accomplish this work knows they have a special place in my heart.

To both my promotors a big big thank you for all the time invested in me. It has been a difficult journey, specially towards the end. Throughout my working days at the EMC you were both there to support me, and when one did not have the time the other did. In that, I think it is a blessing to have two promoters, as participating as you were, with regards to my PhD trajectory. You have witnessed some of the worst moments I have had to overcome at the personal level. At those times, your understanding was beyond this world. But also, you were there for some of the best moments…I got married…I got Violet. In the end, it was your concern for me, not only at a professional but also at a personal level, the best experience of my PhD (besides getting papers accepted in good journals!).

To Kimberley, my paranymph – thank you so much for always being there for me and for having helped so much during this last phase of the PhD. What I appreciate the most and the least about you?! – your brutal honesty ☺ (whether is most or least depends on the kind of day I am having). Most importantly, our friendship remains and will remain even when the PhD is finished.

To Thomas, my other paranymph and previous master student. You were a great help in a difficult time of my PhD and working together we were able to accomplish two great publications. Unfortunately the distance does not make it easy to keep contact but I hope we will stay in touch to share the awesome family moments we will be having ☺

To René, Andrew, Saskia and Elizabeth, the best bioinformatics ever! Thank you so much for all your help with the huge amounts of data I had to deal with.

To Thanassis, Alexandra, Carmen, Gina, Rochelle, Kostas, Katharina, Babis, Rute and Carolina, the closest thing to a family I had in this country for a long time, thanks for being there, both in good and bad moments. I hope you all know how important you are for me! Especially Thanassis – a bit of this thesis is also for you!

Para a minha familia, mãe, pai e Suclas. Infelizmente, tive de fazer esta tese longe de vós, mas Portugal não me deixou grandes alternativas. Obrigada por tudo. Suclas, tu foste certamente o motivo pelo qual vir para a Holanda fazer o doutoramento mais me custou. E serias o único que me faria ter ficado em Portugal se tivesse havido alternativa. Sempre quis ver o mundo, para além do nosso cantinho solarento a beira do Oceano Atlântico, mas nunca pensei ter de pagar um preço tão alto por isso! Deixar-te ficar tão pequenina em Portugal e não estar presente para te ver crescer é das poucas coisas que me arrependo! Espero que um dia possamos voltar a morar no mesmo País. Amo-te!

Para a minha avó Lurdes, já falecida, esta tese é em tua homenagem! Voltar a ver-te é a melhor razão que encontro para acreditar, que algures, existe um céu...

Para os meus amigos em Portugal, obrigada pelos bons momentos, de cada vez que retornei ao caís...foi sempre um prazer ver-vos e pôr a conversa em dia. Espero que, se algum dia voltar (e caramba – espero bem que sim), possamos retomar de onde o tempo parou – no dia em que parti e estar ainda mais presentes na vida uns dos outros!

To my Dutch family, Oma, Ma, Pa, Priscilla and Sergio thank you so much for having taken me in as one of your own and to making me feel at home. What started out difficult, with all the barriers of language and culture, is now what we can call a family relationship, with all its pros and cons. I hope it will keep getting better for as long as we are all here ☺!

To my new friends and colleagues from Groningen. A big big thanks for having received me so well in the group. But most of all, Kirril and Elly who have made this last PhD journey go from tears of sadness to joy – I can definitely say…it was easier with you there for me! Thanks for always listening…

And then, in the end, the biggest thank you of all to my dear husband Jeffrey and my sweet daughter Violet. No words could ever be good or sufficient enough to describe my gratitude towards you…Therefore I will not waste time trying to get close to what is ineffable…You mean all the possible worlds to me! Let a new adventure begin…